CRC 457 / March 1983

ORIGINAL SCALING OF ASVAB FORMS 5/6/7: WHAT WENT WRONG

Milton H. Maier
Ann R. Truss

CENTER FOR NAVAL ANALYSES

Work conducted under contract N00014-76-C-0001

CENTER FOR NAVAL ANALYSES
2000 North Beauregard Street, Post Office Box 11280, Alexandria, Virginia 22311 (703) 998-3500

13 April 1983

MEMORANDUM FOR DSITRIBUTION LIST

Subj: Center for Naval Analyses Research Contribution 457

Encl: (1) CRC 457, "Original Scaling of ASVAB 5/6/7: What Went Wrong," by Milton M. Maier and Ann R. Truss, March 1983

1. Enclosure (1) is forwarded as a matter of possible interest.

2. Research Contributions are distributed for their potential value in other studies and analyses. They do not necessarily represent the opinion of the U.S. Marine Corps or the Department of the Navy.

CHRISTOPHER JEHN
Director
Marine Corps Operations
Analysis Group

Subj: Center for Naval Analyses Research Contribution 457

A1	Assistant Secretary of the Navy (M&RA)
A1	Deputy Assistant Secretary of the Navy (Manpower)
A2A	Comptroller of the Navy
A2A	Office of Program Appraisal
A2A	Chief of Naval Research
A6	Deputy Chief of Staff (Manpower) HqMC
A6	Deputy Chief of Staff (Training) HqMC
	Director, Personnel Procurement Division (HqMC)
	Director, Manpower Plans & Policy Division (HqMC)
	Director, Personnel Management Division (HqMC)
A6	Deputy Chief of Staff (RD&S) HqMC
A6	Fiscal Director, HqMC
A6	Director, History and Museums, HqMC
B1B	Under Secretary of Defense for Research and Engineering
B1B	Assistant Secretary of Defense (MRA&L)
B1B	Office of Ass't Sec'y of Defense (MRA&L)(MP&FM)(AP)
B1B	Director, Program Analysis and Evaluation, OSD
B2A	Defense Technical Information Center (12 copies)
B3	National Defense University
B3	Armed Forces Staff College
FF18	Naval Tactical Support Activity
FF38	Naval Academy (Nimitz Library)
FF44	Naval War College
FF48	Naval Human Resources Management Center
FJ18	Naval Military Personnel Command
FJ76	Naval Recruiting Command
FJ89	Naval Manpower Material Analysis Center, Atlantic
FJ89	Naval Manpower Material Analysis Center, Pacific
FKA6A16	Naval Personnel R&D Center (POM Program Director, Code 12)
	Naval Personnel R&D Center (Technical Library)
FT1	Chief of Naval Education and Training
FT73	Naval Postgraduate School
FT87	Human Resources Management School
V8	Marine Corps Recruit Depot, Parris Island
V8	Marine Corps Recruit Depot, San Diego
V12	Marine Corps Development and Education Command

OpNav:	Op-09BH (Naval History)
	Op-96 (Systems Analysis Division)
	Op-01 (DCNO, Manpower, Personnel & Training)
	Op-11 (Total Force Planning Division)
	Op-13 (Military Personnel and Policy Division)
	Op-15 (Human Resources Management Division)

Subj: Center for Naval Analyses Research Contribution 457

Other

Department of the Army (Attn: Adj Gen'l) (6 copies)
Department of the Army Library
Department of the Army Headquarters (Code DAPE-MPE-CS)
Army Research Institute (Director, Manpower & Personnel Laboratory)
Army Research Institute (Chief, Personnel Utilization Technical Area)
Army Research Institute (Technical Library)
Department of the Air Force (SAMI)
Hq Air Force Manpower & Pesonnel Center (Code MPC/YPT)
Air Force Human Resources Laboratory (AFHRL/MOA)
Air Force Human Resources Laboratory (AFHRL/MOAE)
Air Force Human Resources Laboratory (AFHRL/Technical Library)
Hq, Military Enlistment Processing Command (Code MEPCT-P)
Hq, U.S. Coast Guard (Code G-P-1/2/TP42)
Institute for Defense Analyses
Human Resources Research Organization
The Rand Corporation

CRC 457 / March 1983

ORIGINAL SCALING OF ASVAB FORMS 5/6/7: WHAT WENT WRONG

Milton H. Maier
Ann R. Truss

Marine Corps Operations Analysis Group

CENTER FOR NAVAL ANALYSES

2000 North Beauregard Street, Alexandria, Virginia 22311

ABSTRACT

By April 1976, 4 months after it was introduced, the traditional meaning of scores on the Armed Services Vocational Aptitude Battery, forms 5, 6, and 7 (ASVAB 5/6/7), was being questioned. Scores were found to be too high compared with the traditional reference of the ASVAB score scale in the above-average range. By 1980, scores were also verified to be too high in the below-average range. Our analysis to find the errors in the score scale suggested three reasons:

- Incorrect scoring of the reference test used with the sample of Navy and Air Force recruits
- Coaching on the reference test used for Army examinees
- Using operational test scores as the reference variable for Army examinees and excluding those who did not qualify for enlistment from the calibration sample.

The explanations accounted for almost all the inflated scores on the original scale for ASVAB 5/6/7 above a percentile score of 50 and below a percentile score of 15. Between percentile scores of 15 and 50, however, a residual of up to one-third the difference between the original scale and the traditional ASVAB remained unexplained.

On 1 October 1980, a correct score scale for ASVAB 5/6/7, accurately calibrated to the traditional reference, was implemented.

Based on our analysis we conclude that the original 1976 ASVAB 5/6/7 score scale was in error and that the traditional meaning of the ASVAB scores has been restored.

EXECUTIVE SUMMARY

INTRODUCTION

The Armed Services Vocational Aptitude Battery (ASVAB) is used by the military services to select and classify enlisted personnel. The Armed Forces Qualification Test (AFQT), derived from the ASVAB, is used by the Department of Defense (DoD) to report the mental ability of recruits to Congress. Since it was first introduced in 1950, the AFQT has also been used to track historically the mental ability of recruits. These uses of the ASVAB require a stable score scale that does not change meaning when new versions of the test are introduced.

PROBLEM

When forms 5, 6, and 7 of the ASVAB (ASVAB 5/6/7) were introduced on 1 January 1976, the nominal quality of enlisted recruits immediately increased. The ASVAB scores were too high, and the scale was corrected in the summer of 1976 in the average and above-average range of the scale, but not in the low range, by the ASVAB Working Group. The ASVAB Working Group consists of policy and technical representatives from all services and the Office of the Secretary of Defense (OSD); it has the responsibility to develop and maintain the ASVAB. By 1980, the ASVAB scores were also shown to be too high in the low range. Because the percentage of recruits with low scores did not decrease as dramatically as those in the upper range, some DoD personnel managers questioned whether scores in the low range of the scale were in fact too high.

The differences between the original ASVAB score scale and the correct scale are large. According to the original scale, about 5 percent of DoD enlisted accessions from 1976 until 1980 were in AFQT category IV (percentile scores 10 through 30). According to the correct scale, the number of accessions in AFQT category IV during this period was about 30 percent. The 25 percent difference was upsetting to personnel managers. Some of them expressed concern about the accuracy of the corrections to the ASVAB 5/6/7 score scale.

The purpose of this report is to address the concerns of personnel managers about the accuracy of the ASVAB score scale. The accuracy of the scale is defined in terms of meaning of the scores: the same score should indicate the same level of expected performance regardless of which version of the test is administered. We reanalyzed the sample of examinees, tested in fall 1975, used to construct the original scale for ASVAB 5/6/7. We attempted to determine the causes of the error in the scale and then to verify that the corrections to the scale did in fact restore the traditional meaning of the ASVAB scores.

RESULTS

We found three likely explanations for the inflated scores in the original ASVAB 5/6/7 scale.

- The reference test used with the sample of Navy and Air Force recruits most likely was not scored correctly. (In the Discussion section we provide more information about each of these explanations.)

- The reference test scores for the Army examinees were inflated by coaching on the test.

- Many Army examinees tested at Armed Forces Examining and Entrance Stations (AFEES) appear to have been excluded from the calibration sample, used to construct the score scale, on the basis of their reference test scores. Those who failed to qualify for enlistment (AFQT scores 1 to 15) tended to be excluded from the calibration sample.

When we adjusted the original ASVAB 5/6/7 scale for the effects of these three sources of error, we essentially reproduced the correct scale above a percentile score of 50 and below a percentile score of 15. In the range of percentile scores 15 to 50, however, a residual of up to one-third of the difference between the original and correct scales remained unaccounted for. The original 1976 ASVAB 5/6/7 scale, the correct scale adopted in 1980, and the scale based on our reanalysis are shown in figure I.

DISCUSSION

The research design for calibrating ASVAB 5/6/7 to the traditional score scale was complex. The ASVAB Working Group attempted to minimize the extra testing of applicants for enlistment at AFEES. The upper range of the score scale was to be based on samples of Navy and Air Force recruits tested with an earlier version of the ASVAB (ASVAB 2) as the reference test and the new tests, forms 5, 6, and 7. The low range of the scale was to be based on a sample of Army applicants for enlistment tested at AFEES. The Army applicants were administered the new tests, but not a special reference test. Instead, their enlistment test scores were used as the reference variable for calibrating the new tests. If scores on the reference test are too high, then scores on the new test are also too high. The reference test scores can be too high because they are not scored properly, as we suspect for the Navy and Air Force samples, or because people with low scores on the reference test were systematically excluded from taking the new test, as we suspect for the Army sample, or because people were coached on the reference test. These causes are independent of each other, and their effects are cumulative.

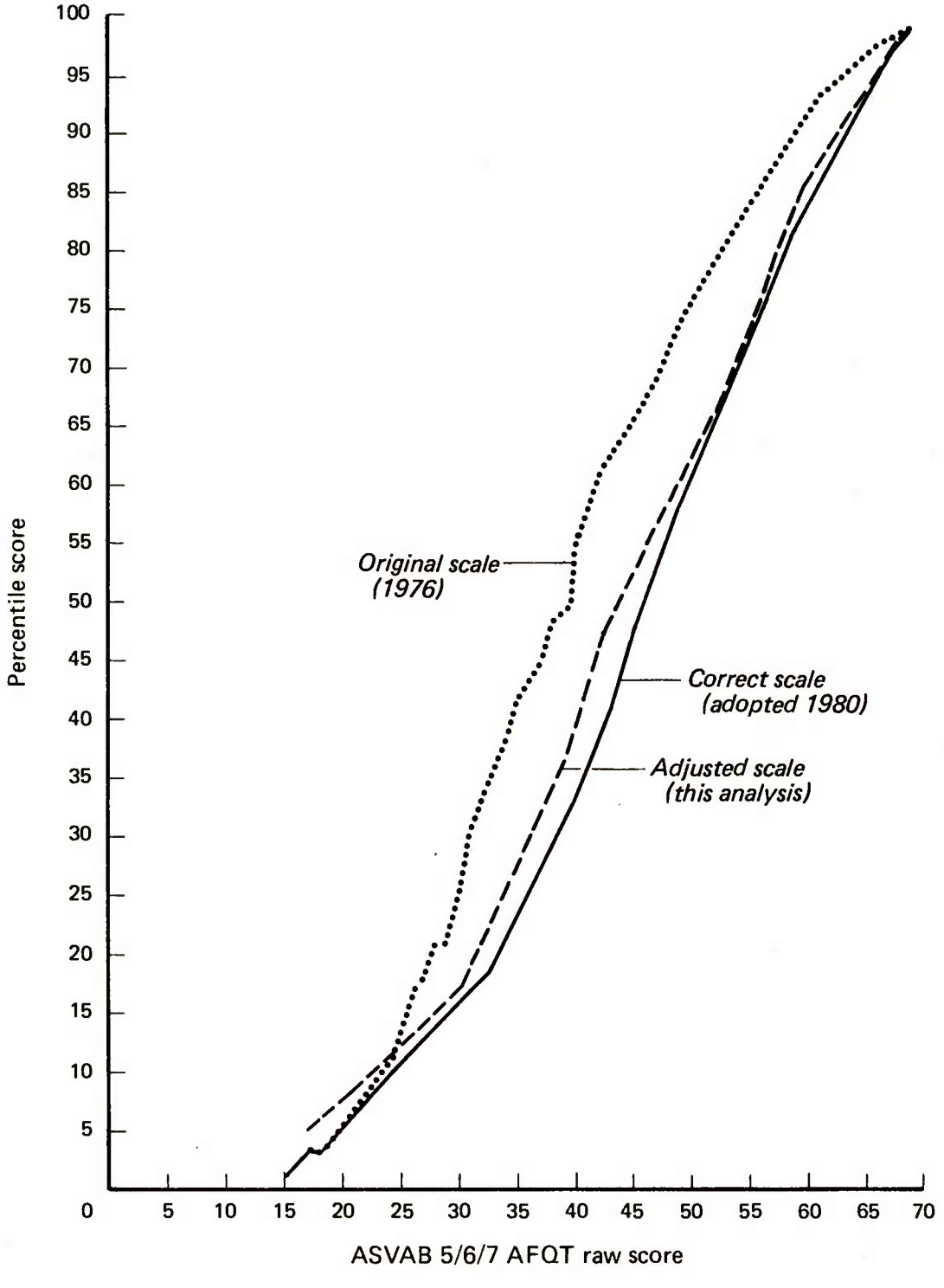

FIG. I: ADJUSTED ASVAB 5/6/7 SCALE COMPARED TO ORIGINAL AND CORRECT SCALES

A scoring error on ASVAB 2 was suggested because the AFQT scores of the Navy and Air Force recruits in the calibration were excessively high compared to all recruits for these services in FY 1975. The proper scoring procedure for ASVAB 2 was to subtract one-third the number of wrong answers from the number of right answers. When we recomputed the ASVAB 2 scores using the proper scoring formula, the scores of the Navy and Air Force samples resembled those of the FY 1975 accessions. Using the rescored reference test, we closely reproduced the correct scale in the above-average range.

For the Army sample, the use of enlistment test scores as the reference variable accounted for about two-thirds of the inflation in the original scale between percentile scores 15 and 50. Use of enlistment tests as the reference variable inflated the original ASVAB 5/6/7 scale because many examinees were coached on them and because many examinees tried harder on them than on the new versions. We estimated the amount of coaching on the enlistment tests for the Army sample by computing a Psuedo AFQT, composed of ASVAB subtests highly correlated with the AFQT, but less subject to coaching, and using it to estimate the amount of coaching on the AFQT. The Pseudo AFQT placed 1.7 times as many Army examinees in AFQT categories IVB, IVC, and V (percentile scores 1 through 20) as did the AFQT, and correspondingly fewer in AFQT category IIIB (percentile scores 31 to 49). We adjusted the distribution of enlistment test scores for the Army examinees by removing the effects of coaching on the tests. We also adjusted the original ASVAB 5/6/7 scale for the estimated effects of using the enlistment tests as the reference. Examinees tend to try harder on tests that affect their qualification for enlistment than on the new versions that do not affect them. The reference test scores therefore would be relatively higher than their scores on the new tests. In addition, many of the failures on the enlistment tests were excluded from the calibration sample, which also results in relatively high reference test scores. After applying these adjustments, a small residual inflation between percentile scores 15 and 50 still remained.

Three possible explanations for the unexplained residual inflation are:

- The correct scale is too difficult in the percentile score range 15 to 50
- The score scales for the reference tests used to calibrate ASVAB 5/6/7 (ACB-73 and ASVAB 2) were themselves inflated
- Other factors were operating to change the meaning of the scale.

The score scales for reference tests used to calibrate ASVAB 5/6/7 appear to have themselves been inflated in the category IVA range. The data available do not permit a precise estimate of the amount of

inflation in their scales. Other factors, including changes in test content, the population, and test calibration procedures, may also help explain the residual inflation. Because of uncertainties in the scores, we could not arrive at any definite explanations for the residual inflation.

The evidence is strong that the correct scale is, in fact, accurate. In spring 1976, the Navy and Air Force checked the calibration of ASVAB 5/6/7 and their results agreed with the correct scale from the top of the scale down to a percentile score of 20; below a score of 20, the data were too scanty for reliable calibration. In 1979 and 1980, three independent studies, based on applicants for enlistment tested at AFEES, Marine Corps recruits tested at reception centers, and high school students tested in geographically dispersed schools, were in essential agreement, and led to DoD's adoption of the correct scale in 1980.

A persistent concern of personnel managers has been why the inflated ASVAB 5/6/7 scale was not apparent in 1976 at the low end of the scale (AFQT category IV). We examined score distributions of Army and Marine Corps applicants when new versions of enlistment tests were introduced during the early 1970s. The results indicate that there is a great deal of uncertainty about the meaning of any scores during this period. The main reason is that coaching appears to have been widespread, which inflated the scores. A second reason is that the scales for some of the tests used during this period may themselves have been inflated. A third reason is that scoring errors may have inflated some of the scores. The effect was to inflate the test scores, which then helped mask the inflation of ASVAB 5/6/7 scores.

LESSONS LEARNED

The main lesson is that the construction of score scales for enlistment tests should be done right in the first place. Even though processing at AFEES is disrupted by the extra testing, the cost is small compared to enlisting large numbers of people who should not qualify. A related lesson is that test scores used to determine qualification for enlistment should not also be used as the reference variable for calibrating new versions. The reference test and new versions should always be given together in a separate testing session under identical testing conditions. The third lesson is that all enlistment test scores during the 1970s have a large degree of uncertainty, and extra caution is required when drawing conclusions about the mental aptitudes of recruits.

CONCLUSIONS

Our conclusions are that the original ASVAB 5/6/7 score scale was inflated throughout the score range and that the traditional meaning of the ASVAB score scale has been restored.

TABLE OF CONTENTS

TABLE OF CONTENTS (Cont'd)

TABLE OF CONTENTS (Cont'd)

LIST OF ILLUSTRATIONS

LIST OF ILLUSTRATIONS (Cont'd)

LIST OF TABLES

LIST OF TABLES (Cont'd)

CHAPTER 1

INTRODUCTION

BACKGROUND

The Armed Services Vocational Aptitude Battery (ASVAB) is used by the military services to select and classify enlisted personnel. The ASVAB was developed to predict performance in the military--people with higher scores should perform better than those with lower scores. The ASVAB is also used by the Department of Defense (DoD) to report the ability or quality of enlisted accessions to Congress. Another major use of the test is to follow historically the ability of enlisted personnel for each service and for DoD as a whole. Overall, the ASVAB plays a vital role in the management of military personnel.

The usefulness of the ASVAB to personnel managers is greatly enhanced by a stable score scale that does not change meaning when new versions of the test are introduced or when the general ability of recruits changes, as between a draft and all volunteer environment. The meaning of the scores, in terms of performance expected from people with the same test scores, should remain relatively invariant across time. Then personnel managers can be reasonably confident that decisions made on the basis of ASVAB scores will have the intended effects. For example, when enlistment standards are raised or lowered, the level of performance should change correspondingly. If the meaning of the ASVAB scores changes, and managers do not know what the new scores mean relative to their traditional expectations, then personnel decisions may not have their intended effects.

The AFQT

The Armed Forces Qualification Test (AFQT), derived from the ASVAB, is widely used by personnel managers. The first AFQT was introduced in 1950. Since then it has been used for the three major purposes of ASVAB mentioned: to develop the first screen for selecting recruits; to report the quality of recruits to Congress; and to track historically the quality of recruits. The ASVAB also provides aptitude composite scores used to classify recruits to their skill training courses. Our focus in this report is on the AFQT because it is so widely used and because whatever we learn about the AFQT is also true about the aptitude composites.

Traditionally the AFQT score scale has been divided into five categories (table 1) for reporting quality of enlisted accessions. In general, commissioned officers and most noncommissioned officers score in AFQT categories I and II. The middle category, III, is further divided into two subcategories: IIIA, percentile scores 50 through 64; and IIIB, percentile scores 31 through 49. Category IV is the range in

which minimum enlistment standards usually have been set. It is further divided into three groups: IVA, percentile scores 21 through 30; IVB, 16 through 20; and IVC, 10 through 15. No service currently enlists anyone in category IVC, and as a rule only high school graduates are accepted in categories IVA and IVB. Nongraduates must be in category III or above. Persons in category V have not been accepted for enlistment or induction since World War II.

TABLE 1

ARMED FORCES QUALIFICATION TEST (AFQT) CATEGORIES

AFQT category	Percentile score range	Percent of reference population	Description
I	93-100	8	Above average
II	65-92	28	Above average
III	31-64	35	Average
IV	10-30	21	Below average
V	1-9	9	Unqualified for service

How to Maintain a Stable Score Scale

The procedure for maintaining a constant score scale is to calibrate new versions of the test to a reference test. The score scale for the reference test is already known to be accurate. The ASVAB and previous versions of military selection and classification tests have been referenced to the scale of the Army General Classification Test (AGCT), widely used during World War II. The meaning of the military selection and classification tests had remained relatively constant because they were referenced to the same scale. When forms 5, 6, and 7 of the ASVAB (ASVAB 5/6/7) were calibrated in fall 1975, however, an error was made.

The proper calibration procedures require that the reference test scores be accurate, that both the reference and new test (in this case ASVAB 5/6/7) be administered under the same testing conditions, and that the analysis be performed correctly. Errors from these sources will distort the score scale for the new test, and the traditional meaning of the scores will be lost.

PROBLEM

The accuracy of the score scale for ASVAB 5/6/7 was questioned within 4 months after its introduction on 1 January 1976. The percentage of applicants for enlistment with above-average scores (AFQT categories I and II) increased immediately after the battery was introduced. For Army male applicants in categories I and II the increase was from 20 to 28 percent.

In spring 1976, the ASVAB Working Group, composed of technical and policy representatives from all services and the Office of the Secretary of Defense, initiated studies to check on the magnitude of the scaling error. The ASVAB Working Group is responsible for developing and maintaining the ASVAB. Data were collected on Navy, Air Force, and Marine recruits, and the results consistently showed that the scores were inflated. Inflation means that the scores indicate a higher level of expected performance than they should compared to their traditional meaning. The Working Group lowered scores in the average and above-average range (AFQT categories I, II, and III) in fall 1976, but they left scores in the lower range (categories IV and V) virtually unchanged. The studies in 1976 that first confirmed a problem with the ASVAB score scale are described in appendix A.

Concern about the accuracy of the scale persisted in the ASVAB Working Group. By spring 1980, three independent studies confirmed that the scale was seriously inflated in the below-average range (AFQT category IV) [1, 2, 3]. The Office of the Secretary of Defense retained three testing consultants to review the results of the three studies [4]. They agreed that the scale for ASVAB 5/6/7 produced by these studies was more accurate than the scale being used, and recommended that the new scale, called the "correct scale" in DoD, should replace the existing one.

ASVAB 5/6/7 Score Scales

By 1980, three ASVAB 5/6/7 score scales had been developed and used:

- Original Scale--adopted 1 January 1976 with the initial introduction of ASVAB 5/6/7, and used until September 1976 to select and classify enlisted recruits
- Operational Scale--adopted September 1976, and used until 1 October 1980 to select and classify enlisted recruits
- Correct Scale--adopted in 1980, and used to report retrospectively the quality of enlisted recruits from 1 January 1976 until 1 October 1980.

The correct scale was not used to select and classify recruits. ASVAB 5/6/7 was replaced with new versions on 1 October 1980 (forms 8, 9, and 10). The new versions were accurately calibrated to the reference test, and the traditional meaning of selection and classification decisions based on ASVAB scores was restored.

Constructing the AFQT 5/6/7 Score Scale

The AFQT from ASVAB 5/6/7 contains 70 items, which means it has 71 raw score points (0 through 70). The raw scores by themselves have no meaning. They acquire meaning, or can be interpreted, when converted to the score scale of a reference test administered to a defined population. The process of converting raw scores to percentile scores on reference tests is called scaling, calibration, or, sometimes, normalization. See appendix B for a more complete discussion of procedures for scaling a test and interpreting ASVAB scores.

The AFQT score scale is expressed as percentile scores, where the basis for computing percentile scores is the World War II (WWII) mobilization population. Each of three ASVAB score scales produced a different conversion from AFQT raw score to percentile score (figure 1). Each conversion in figure 1 purports to be referenced to the WWII population and thereby conveys the traditional meaning of the scores. Yet the original scale and correct scale differ by up to 23 percentile score points (at raw score 41).

Differences of this magnitude have a severe effect on the nominal quality of enlisted accessions. According to the operational scale (used to report the quality of enlisted accessions from September 1976 until October 1980), only about 5 percent of the enlisted accessions were placed in category IV. According to the correct scale (used in 1980 to recompute the quality of enlisted accessions) the percentage should have been 30, an increase of 25 percentage points. Up to one-quarter of all recruits accessioned while ASVAB 5/6/7 was in use would not have qualified for enlistment if the correct scale had been used.

While ASVAB 5/6/7 was used to select and classify recruits, personnel managers thought they were making decisions based on the traditional meaning of ASVAB scores. When they were told that the quality of accessions was much lower than had been reported, many were incredulous. Some managers questioned whether the "correct scale" was, in fact, accurately referenced to the traditional scale. They also wanted to know how such a large calibration error could have occurred in the original scale, and if it did occur, why the inflated scale at the low range of scores was not detected when ASVAB 5/6/7 was first introduced. In this report, we address these concerns of personnel managers.

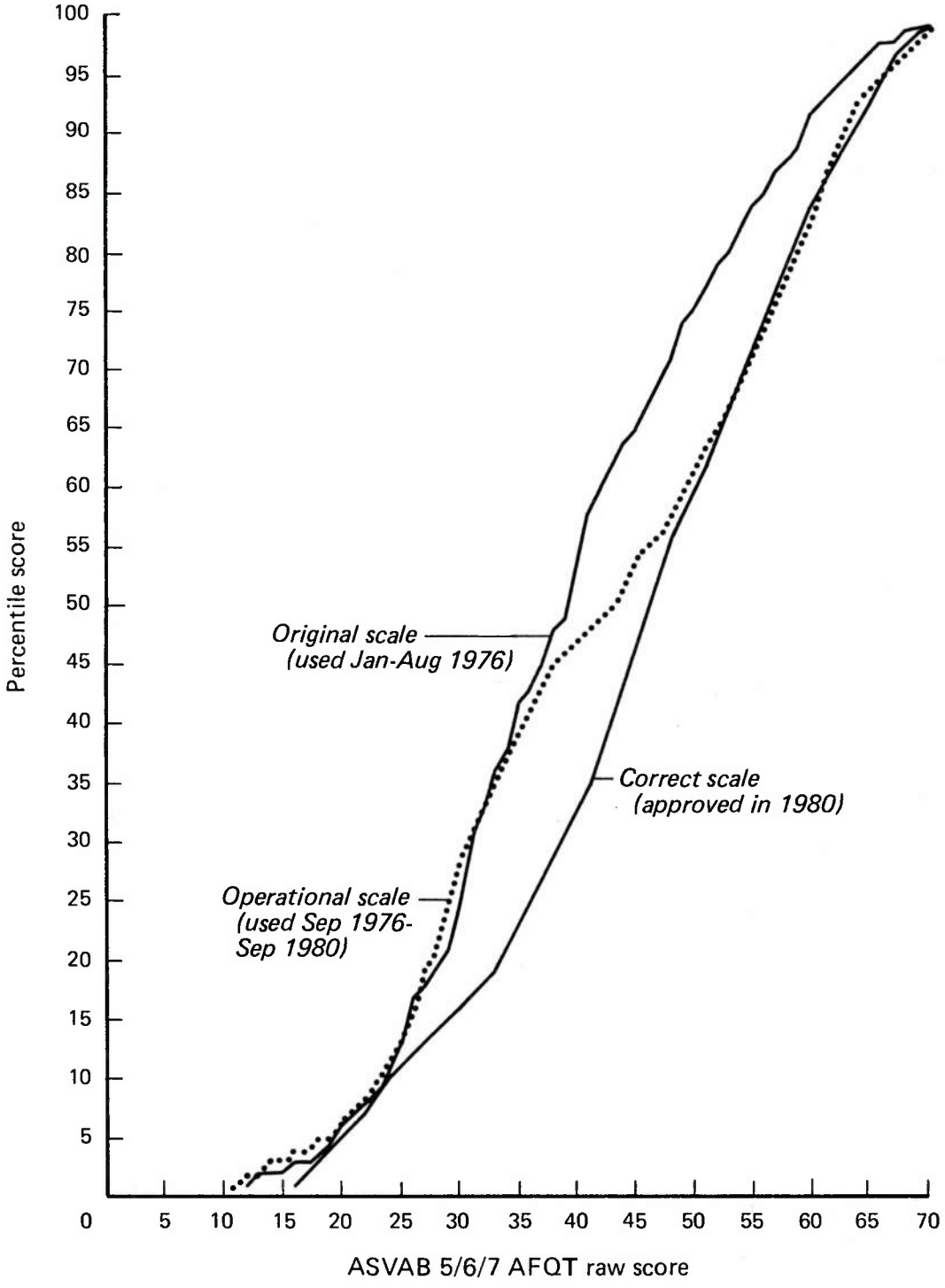

FIG. 1: CONVERSION OF ASVAB 5/6/7 AFQT RAW SCORE TO PERCENTILE SCORE FOR THREE ALTERNATIVE SCORE SCALES

PURPOSE OF ANALYSIS

The purposes of this analysis are to:

- Determine what went wrong with the original scaling of ASVAB 5/6/7
- Reproduce the correct scale in the sample of examinees used for constructing the original scale
- Discuss problems of interpreting score distributions during the All Volunteer Force (AVF) era.

To accomplish the first two purposes, we reanalyzed the data from the sample used for constructing the original scale. We also obtained data from the Defense Manpower Data Center (DMDC) to use in our analysis.* Our analysis focused on the three requirements for a correct calibration that we mentioned earlier in the Introduction:

- Computational accuracy
- Accurate reference test scores
- Proper testing conditions.

We first checked the computations to make sure that simple errors did not explain the inflated scale. Then we checked the accuracy of the reference test scores. Finally, we examined the testing conditions to evaluate their affect on the scores.

Our analysis is perforce retrospective. We are unable to prove that events transpired as we conclude. All we can do is build a plausible argument by showing that our explanations account for the inflation. The original calibration data were collected in fall 1975. At this late date there is no way to reconstruct what the research team actually did when they developed the original scale or how the tests were actually administered to the examinees. We must remain content with a reasonable set of explanations.

ORGANIZATION OF REPORT

In chapter 2 we present our reanalysis of the data from the original calibration sample. We present plausible explanations for why the inflation occurred.

* The support of DMDC in providing the data was invaluable for completing the analysis; Mr. Les Willis of DMDC was especially helpful.

In chapter 3 we attempt to reproduce the correct scale in the original calibration sample. Our explanations from chapter 2 did not account for all the inflation of the scale; therefore, we consider possible explanations for the residual inflation.

In chapter 4 we discuss problems interpreting score distributions during the AVF era. The awareness of these problems grew out of our search for the errors in the miscalibration of ASVAB 5/6/7.

CHAPTER 2

SOLVING THE PROBLEM

INTRODUCTION

Because the original ASVAB 5/6/7 score scale was in error the first step was to locate the sources of the error and then to estimate their effects on the inflated scale. To do that we needed to reanalyze the data from which the original score scale was developed. Data for the original sample was provided by the Air Force Human Resources Laboratory (AFHRL), the executive agent for research on the ASVAB. In the remainder of the this subsection, we review the research design for developing the original score scale. Our review focuses on likely sources of error: computational or clerical-type errors, or errors that make the reference test scores too high or the experimental (ASVAB 5/6/7) scores too low.

Research Design

All data collection took place in fall 1975 [5]. Navy and Air Force recruits were tested at reception centers in special testing sessions. Because Navy and Air Force recruits tend to be above average in ability, they were used to establish the top half of the scale. The bottom half of the scale was established on a sample of Army applicants for enlistment, who tend to be below average in ability. Army applicants were tested at Armed Forces Examining and Entrance Stations (AFEES).* Results for these samples were combined when the original scale was computed. Over 5,000 examinees were tested in fall 1975. The sample was reduced to 1,600 cases, and the original scale was computed on the 1,600 cases [6]. Our first check will be to recompute the scale in the sample of 1,600 to see if simple computational errors occurred. Then we check to see if reducing the sample from 5,000 to 1,600 cases introduced bias that could have inflated the scale. We also checked the scoring of experimental answer sheets. The examinees coded the form number of the test (5, 6, or 7) they were taking on their answer sheets, and this number governed the scoring key used with their answer sheet. If an examinee coded the wrong number, the wrong key would be applied and the score could be much too low. In addition, the keys themselves could have been in error. When the data were originally analyzed, the keys were still experimental. Given the pressure to complete the analysis in time for introduction of the tests on 1 January 1976, there

* The name for AFEES was changed to Military Entrance Processing Stations (MEPS) in 1982. We used the earlier name because that is what they were called when the data were collected.

was little time to exercise adequate quality control, and clerical-type errors could have easily occurred.

Navy and Air Force Recruits

Because the testing procedures for the sample of Navy and Air Force recruits were different from the sample of Army examinees, we analyzed them separately. The Navy and Air Force recruits were administered the reference test (ASVAB 2, an earlier version used in the high school testing program) and the experimental tests in special testing sessions. All testing was done at reception centers by trained personnel. We examined the reference test scores to see if they were too high compared to other Navy and Air Force recruits accessioned during the same time period. Other than using the wrong scoring keys, we had no reason to question the accuracy of their experimental test scores. Any errors we find in this sample will affect scores primarily in the upper half of the scale because relatively few of the recruits score in AFQT category IV.

Army Examinees

The search for errors in the sample of Army examinees was more complex. The key feature of the research design that could lead to an inflated scale is the use of enlistment test scores as the reference variable for calibrating ASVAB 5/6/7. The enlistment test used by the Army at that time was the Army Classification Battery, form 73 (ACB-73). Because the reference test determined qualification for enlistment, while the experimental test was just an extra burden on test administrators and examinees, the testing conditions were different. The most likely result is that the reference test scores were too high and that the experimental test scores were too low.

One reason the reference test scores were too high is that many examinees were coached on the reference test. Another reason is that AFEES personnel are likely to select examinees to take the experimental ASVAB on the basis of their reference test scores. Each AFEES in the study was given a quota of examinees they needed to test. However, only examinees with AFQT scores of 50 and below, obtained from ACB-73, counted toward the quota. Although the AFEES were instructed to administer the enlistment and experimental tests in counterbalanced order, the research design invited them to administer the enlistment test first, and then give the experimental test only to those who score 50 and below on the AFQT.

The AFEES were also given two other instructions that likely influenced their testing procedures:

- Limit testing to 1 day. No examinees were to be held overnight. Participants in the study could have taken 7 hours of testing in 1 day--4 hours for the enlistment test and 3 hours for the experimental test. The extra 3 hours had to be squeezed into an already full day of processing.

- Test, if desired by AFEES, those who had previously taken the enlistment tests. These were individuals reporting to the AFEES for physical examinations only, and those in the Delayed Enlistment Program (DEP).

A likely scenario is that the AFEES would attempt to reduce their extra testing burden by administering the experimental tests to those already qualified for enlistment. To the extent that AFEES followed this practice, failures on the enlistment tests and those who scored above 50 on the AFQT would tend to be excluded from the calibration sample. Because many examinees would know that they had already qualified for enlistment, their motivation to take experimental tests would be lower. Selecting examinees on the basis of their reference test scores, therefore, would result in an inflated score scale.

The opportunities for errors that inflated the ASVAB scale are numerous. To anticipate the results of this chapter, we did not find a simple explanation that accounted for all the inflation. Instead, we laboriously examined the possibilities. The results of our analysis are presented in about the same sequence that we listed the possible sources of error. The general strategy was to start with the most obvious and objective possible explanations, such as computational errors, and proceed to more subtle effects, such as selection of examinees on the basis of reference test scores, only when the more obvious possible explanations did not work.

REANALYSIS OF DATA FOR THE ORIGINAL CALIBRATION SAMPLE

Our reanalysis of the data for the original calibration sample involved two steps. The first step in our analysis was to verify that the arithmetic computations in the sample of 1,600 cases for the

original scale were correct.* The original scale was contained in an official DoD publication [7].

Second, we computed the percentile scores in the full sample of about 5,000 cases. Results of our recomputation in the sample of 1,600 cases essentially agreed with those of the original score scale, except at the lower end. The original scale has irregular conversions from raw score to percentile score. For example, raw scores 28 and 29 convert to a percentile score 21, but a raw score 30 converts to a percentile score 25. The conversion from our reanalysis of the data is much smoother (figure 2). Our recomputation suggests there were no errors in the original arithmetic that helped explain the original scaling error.

Analysis of the Full Sample

Bias might have been introduced when the sample was reduced from about 5,000 to 1,600 cases. In addition, errors may have arisen because the sample of 1,600 contained both males and females, whereas the ASVAB score scale traditionally has been computed only on males. We analyzed the females (N = 558) and males (N = 4,588) separately to determine if the inclusion of females helped explain the inflation of the original scale. Computing a scale for the sample of 4,588 males, then, checked two possible effects: one, that the deletion of about 3,500 cases to stratify the sample may have introduced bias into the score distributions; and two, that the inclusion of females may have affected the score scale. To recompute the score scale on the full sample of 4,588 males we used the equipercentile equating technique (explained in appendix B).

The score scales for the full sample of males, the sample of females, and the calibration sample are almost the same above a percentile score of 40 (figure 3). (The cumulative frequency distributions are in appendix C.) Below that point the scales for the samples of males and females are higher than the original scale. Including females in the original scaling did not introduce bias into the results. The

* The sample was obtained as follows: Each form of ASVAB 5/6/7 was administered to about 1,700 persons. The reference test score distribution for each form of the ASVAB was computed and divided into deciles (10-point intervals). For each form, the smallest number in a decile was used to set the sample size for that form. The sample size for each form was 400 to 600 cases. By randomly deleting cases in each reference test score decile, the number of cases in each decile for each form was set equal. In this way the sample was stratified to have a flat distribution of reference test scores, similar to the traditional reference population.

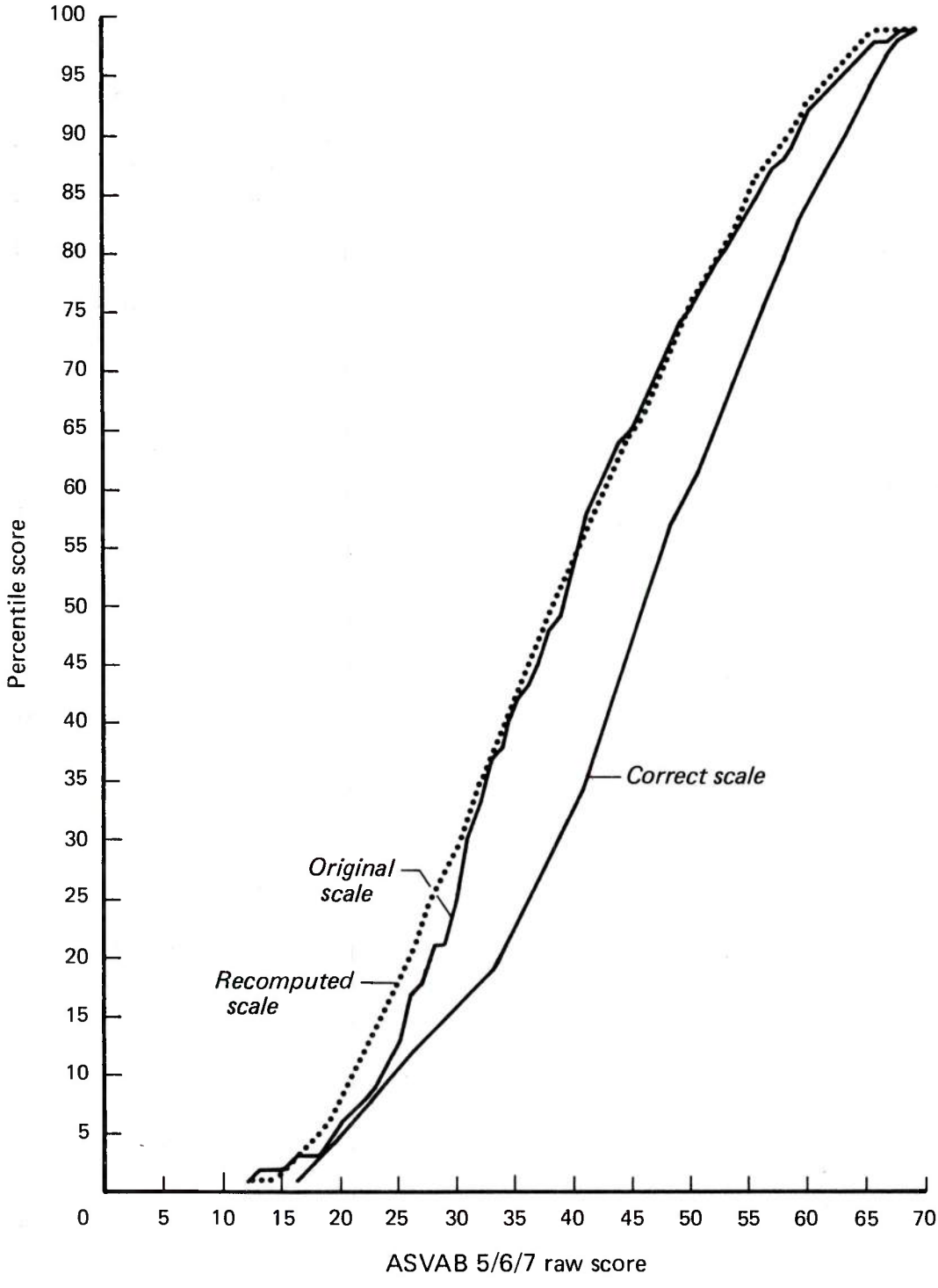

FIG. 2: RECOMPUTATION OF ASVAB 5/6/7 PERCENTILE SCORES IN THE STRATIFIED SAMPLE OF 1,600 CASES

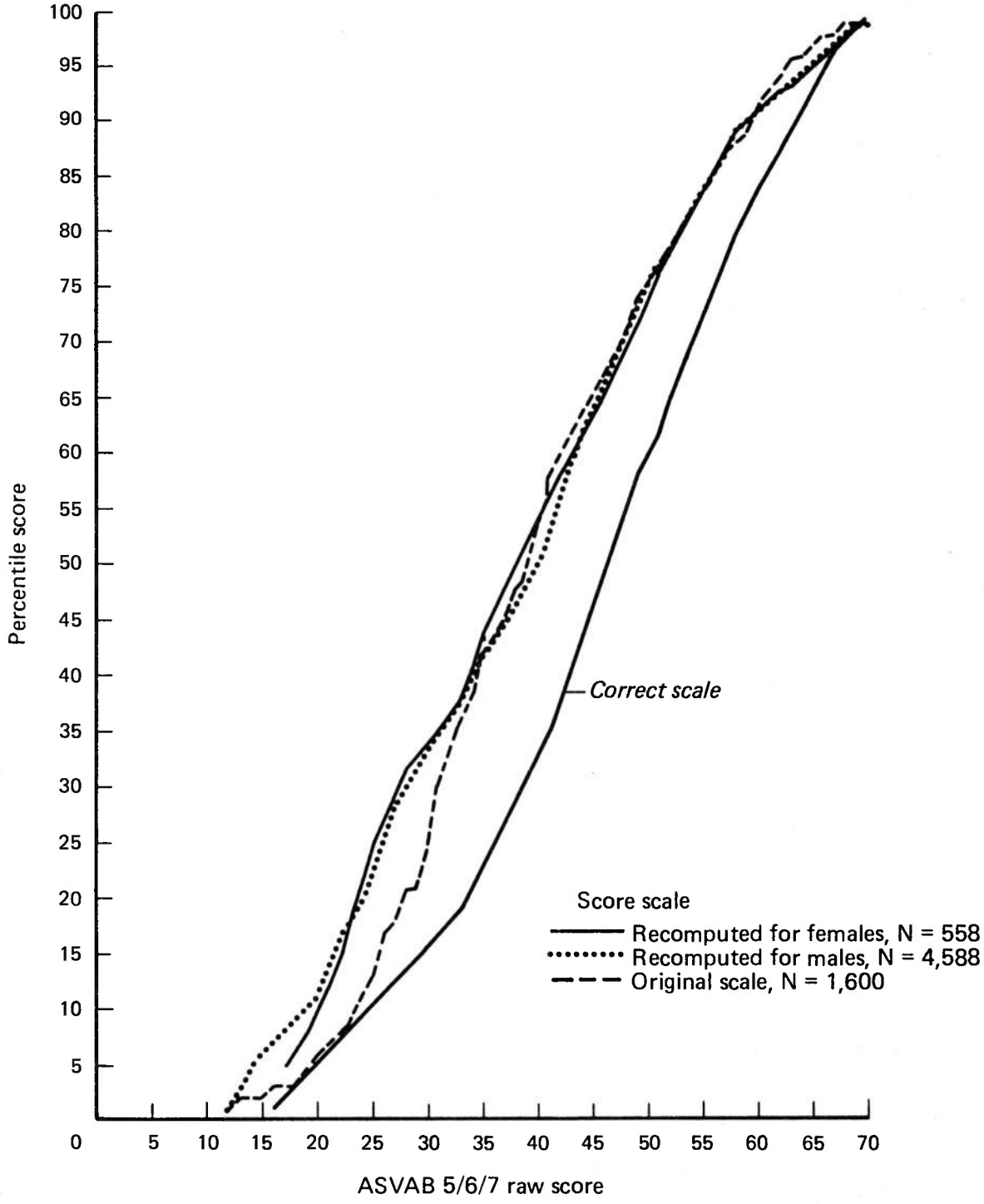

FIG. 3: SCALING OF ASVAB 5/6/7 IN THE FULL ORIGINAL CALIBRATION SAMPLE

results for the full sample agree closely with our recomputation of the sample of 1,600 cases. All these scales based on the original sample are inflated compared to the correct scale.

Our recomputation of the ASVAB score scale in the original calibration sample, however, did not reproduce the original scale. From the data available in the calibration sample there is no way to reproduce the original scale. Our analysis did not reveal any computational errors that explained the inflation of the original scale. In fact, our recomputations showed that based on the data available for the calibration sample, the original score scale was too difficult, rather than too easy, in categories IV and V.

We conclude that neither computational errors, reduction of the sample from more than 5,000 cases to 1,600 cases, nor the inclusion of females in the original calibration explains the error in the original scale.

Are the ASVAB 5/6/7 Scores Too Low?

The ASVAB 5/6/7 scale could be inflated because the experimental ASVAB 5/6/7 raw scores were too low. To determine if that is the case we looked at raw scores.

Incorrect scoring keys may have been applied to ASVAB 5/6/7 answer sheets. In addition, last minute changes were made to the test booklets by pasting easy items at the end of some subtests. If the changes were not made to all test booklets, then some examinees would, in effect, be taking different tests. Incorrect scoring and taking different tests could explain the low scores.

To try to solve the problem we developed an empirical scoring key for each form (5, 6, or 7) of the AFQT subtests (Work Knowledge, Arithmetic Reasoning, and Spatial Perception). More able examinees should select the correct answer more often than incorrect alternatives. We counted the number of examinees in the total sample who selected each alternative. In developing the empirical keys, we counted the most popular response as the correct answer. We rescored the AFQT subtests for the Army examinees using the empirical keys. We rescored the answer sheet for each examinee three times, using the empirical keys for each of the three forms.

The original subtest scores, based on the test form coded by examinees on their answer sheets, agreed almost perfectly with the appropriate empirical key. In only 29 cases, or about 1 percent, was there a clear indication that the wrong form had been coded on the

answer sheet. We conclude that the scoring of ASVAB 5/6/7 was done correctly.*

REANALYSIS OF DATA FOR NAVY AND AIR FORCE RECRUITS

Not able to explain for the inflation in the original scale, we then examined the scores for the Navy and Air Force recruits. We checked to see if the reference test scores were too high. We computed the reference test score distributions for Navy and Air Force recruits (table 2). For comparison, the AFQT distribution of all Navy and Air Force recruits in FY 1975 are also shown. The FY 1975 distributions were obtained from the Defense Manpower Data Center (DMDC). The recruits in the calibration sample had much higher test scores compared to the full set of recruits. For example, 10.9 percent of the Navy recruits in the calibration sample scored in category I, compared to only 2.8 percent of all Navy recruits in FY 1975.

A likely possibility is that the reference test was not scored correctly. The ASVAB 2 raw scores, used as the reference test, are supposed to be corrected for guessing by subtracting one-third the number of wrong answers from the number of items correct; omitted items are not counted in the score. ASVAB 5/6/7 is properly scored simply as the number of correct answers with no correction for guessing. Given the time pressure to complete the original scaling of ASVAB 5/6/7, the correction for guessing in ASVAB 2 could easily have been overlooked.

We rescored the ASVAB 2 scores on the AFHRL tape by assuming that each recruit had attempted all items; the number wrong therefore was the difference between the number of items correct and the total number of items. The rescored ASVAB 2 scores were converted into new percentile scores based on the correct scoring formula. Most examinees were able to finish all ASVAB subtests, except for the speeded subtests. The time limits are set to allow more than 90 percent of examinees in the average range to complete the subtests. Because most of the Navy and Air Force recruits score average or higher on the ASVAB, our assumption that each recruit attempted to answer all questions should be generally true.

The ASVAB 2 score distributions based on recomputed scores are shown in the middle columns of table 2. The corrected score distributions are much closer to the distribution for the full year input than

* We did discover one error in the original key for one form of the Spatial Perception (SP) test. The effect of the error, however, was to increase the SP score of the Army examinees rather than lower them. The alternative keyed as correct originally was more popular among the Army examinees than was the correct answer. This error was corrected in subsequent revisions of the operational scoring key.

are the original score distributions. Some examinees in category IV may not have been able to finish the subtests. To the extent that they omitted items, our scoring procedure, which assumed they attempted all items, would reduce their scores below their correct level.

TABLE 2

DISTRIBUTION OF ASVAB 2 REFERENCE TEST SCORES FOR NAVY AND AIR FORCE RECRUITS IN CALIBRATION SAMPLE COMPARED TO FY 1975 RECRUITS

A. Navy recruits

AFQT category	Percent in AFQT Category: Original calibration sample[a]	Recomputed calibration sample[b]	FY 1975 recruits
I	10.9	5.2	2.8
II	58.8	33.8	35.2
IIIA	21.0	25.3	30.0
IIIB	7.0	24.6	27.2
IV	2.3	11.1	4.8

B. Air Force recruits

AFQT category	Original calibration sample[a]	Recomputed calibration sample[b]	FY 1975 recruits
I	15.8	5.9	3.9
II	66.6	48.1	40.0
IIIA	14.5	25.0	30.3
IIIB	3.1	17.0	25.4
IV	0	4.0	0.4

[a]ASVAB 2 scores used in original calibration of ASVAB 5/6/7
[b]ASVAB 2 scores recomputed by subtracting one-third the number of wrong answers from the number of correct answers

Score scales were computed separately for the Navy and Air Force recruits, using both the initial and rescored reference test scores. The conversion from ASVAB 5/6/7 raw score to percentile score for the sample of Navy recruits is shown in figure 4. The conversion for the Air Force recruits is similar to that for the Navy recruits. For comparison, the correct scale is also shown in figure 4. The cumulative

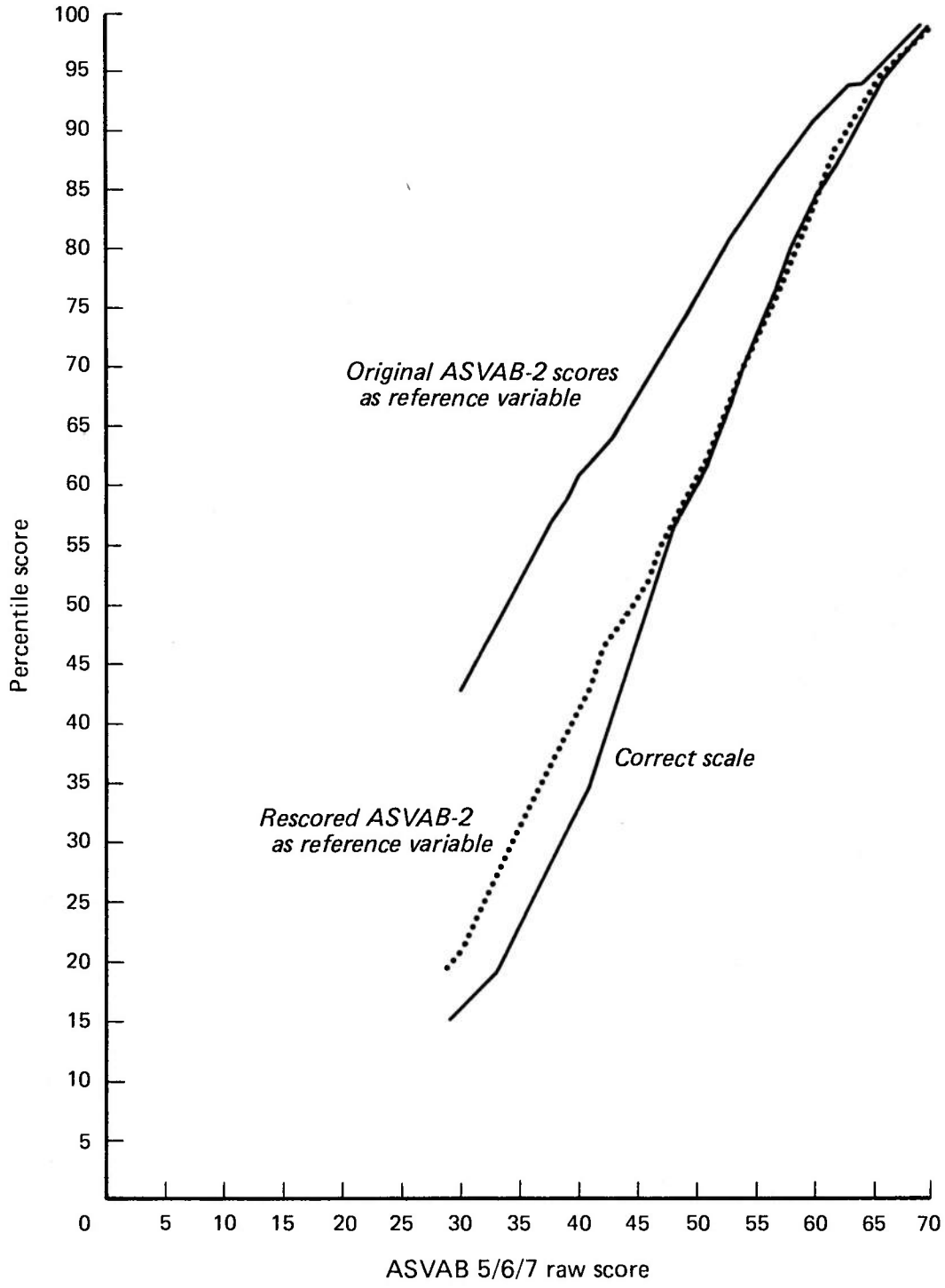

FIG. 4: CALIBRATION OF ASVAB 5/6/7 IN THE SAMPLE OF NAVY RECRUITS

frequencies of the test scores for each service are given in appendix C. The conversions based on the rescored reference test reproduces the correct score scale above a percentile score of 50, where our assumption that the recruits attempted all items is most likely valid.

Our analysis supports our supposition that the ASVAB 2 reference test for Navy and Air Force recruits was scored incorrectly. First, we found that the AFQT distributions based on recomputing the scores agreed more closely with the distributions for the full year's input than did the distributions based on the original scoring. Second, we found that the score scale based on the recomputed ASVAB 2 scores agrees almost perfectly with the correct scale above a percentile score of 50. We conclude, therefore, that the inflation of the original scale in the upper half was a result of incorrect scoring of the reference test for the Navy and Air Force recruit samples.

REANALYSIS OF DATA FOR ARMY EXAMINEES

Prior to reanalyzing the data for the Army examinees, we need to determine the accuracy of ACB-73 score scale, used as the reference test for the ASVAB 5/6/7 scale. If the ACB-73 scale is inflated, then the inevitable result is that the ASVAB 5/6/7 scale is also inflated. After checking the accuracy of ACB-73 scale, we proceed to look for clerical errors. Then we examined the reference test scores to see if they were too high and finally, the experimental ASVAB 5/6/7 scores to see if they were too low.

Accuracy of the ACB-73 Scale

Our conclusion from analyzing the accuracy of the ACB-73 scale is that it is sufficiently accurate. That is, any inflation of the ACB-73 scale does not explain the error in the original ASVAB 5/6/7 scale. Its accuracy is sufficiently questionable, however, that it could explain some of the inflation of the original ASVAB 5/6/7 scale. Because of uncertainties in the data, we cannot estimate accurately the degree to which the ACB-73 scale is inflated. A detailed description of the calibration of ACB-73 is presented in appendix D, together with some checks we computed for this report on the accuracy of the scale.

Clerical Errors

AFEES personnel coded the AFQT score obtained from ACB-73 on the ASVAB 5/6/7 answer sheet for each examinee. Incorrect coding of the AFQT scores could have distorted the ACB-73 score distribution for the Army sample. To check the accuracy of the coding, we obtained the ACB-73 AFQT scores of record from automated files of the Defense Manpower Data Center (DMDC) and compared then to the scores coded on the ASVAB 5/6/7 answer sheets. The mean and standard deviation of the scores of record and the scores coded on the answer sheets are similar; the intercorrelation between the two sets of AFQT scores is high (above

.9). The results suggest there are no large errors in the coded scores. In appendix E, we present details about the accuracy of coding shown by date of testing; we also show the effects of date of testing on inflation of the ASVAB 5/6/7 score scale. In appendix F, we extend the results of appendix E by examining the effects of administering the reference test (ACB-73) before the experimental tests (ASVAB 5/6/7).

We computed the scatterplots between the two sets of scores to check for errors that might have been masked by the summary statistics. There was a tendency to code scores of record that exceeded a percentile score of 50 as either 47 or 50. This systematic error in coding would affect the conversion from raw score to percentile score in the above-average range. In the lower range, however, the coding is either accurate or the errors are random. We concluded, therefore, that errors in coding the ACB-73 scores did not seriously distort the score distributions or inflate the ASVAB 5/6/7 score scale.

Coaching

Many of the AFQT scores in the Army sample may have been inflated because examinees were coached. Before ASVAB 5/6/7 calibration data were collected in fall 1975, ACB-73 had been in operational use for more than 2 years. Recruiters had ample time to become familiar with the test items and develop techniques for coaching examinees. To the extent that examinees were coached on the AFQT, their reference test scores are inflated. The ASVAB 5/6/7 scale would be correspondingly inflated.

To estimate the amount of compromise in the Army sample we needed an independent measure of the examinees' aptitudes. The estimate should be highly correlated with AFQT, but less subject to compromise. The Pseudo AFQT (PAFQT) is such a measure. It consists of other ASVAB subtests with content similar to the AFQT. Since the late 1970s, the PAFQT has been used to help identify suspected compromise on the AFQT.

Pseudo AFQT

The principle underlying the PAFQT is that it is less subject to coaching. Recruiters are more likely to coach examinees on the AFQT subtests than on the other subtests. The PAFQT score is subtracted from the AFQT score. If the AFQT score exceeds the PAFQT by a specified level, then compromise is suspected and the examinee is retested with another form of the AFQT. The second AFQT score becomes the score of record, and it is used to help determine qualification for enlistment.

ACB-73 did not have a PAFQT score; we had to develop one for our analysis. We used the ACB-73 calibration sample to determine the subtests to include in the PAFQT and to calibrate PAFQT to AFQT. (Details of developing the PAFQT for ACB-73 are in appendix G.) The correlation between PAFQT and AFQT was .87 in the ACB-73 calibration sample, which means that PAFQT is a satisfactory estimate of AFQT. We used the

equipercentile equating technique to put PAFQT on the AFQT scale. (The conversion of PAFQT raw scores to AFQT percentile scores is in appendix G.)

PAFQT scores could not be computed for all cases in the Army sample. To compute PAFQT scores, we used ACB-73 subtest scores. The Army did not record ACB-73 subtest scores in personnel records. Fortunately, when the ASVAB 5/6/7 calibration data were collected, ACB-73 answer sheets were collected for 488 Army examinees. We computed PAFQT scores for these 488 cases and used the difference between their AFQT and PAFQT as our basic measure of test compromise.

For Army examinees the difference between AFQT and PAFQT scores may be a conservative estimate of the amount of compromise. Army applicants had to qualify on both the AFQT and on one or two aptitude composites depending on their AFQT score and level of education. Army applicants may have been coached on both the AFQT and on other subtests in ACB-73, including those in PAFQT. To the extent that Army examinees were also coached on the subtests in PAFQT, differences between the AFQT and PAFQT scores are reduced. Even though the estimates of compromise may be conservative, the PAFQT still provides a useful indicator of compromise; together with other estimates we can obtain an AFQT score distribution that is reasonably close to a comparable group of examinees with no test compromise.

Estimated Amount of Compromise in the Army Sample

In the 1975 sample, the estimated amount of compromise is 23 percent. The procedures for estimating the percentage of compromise in a sample are described in appendix G. The percentage of Army examinees in AFQT categories IVB, IVC, and V (percentile scores 1 through 20) should be increased from 12.5, as measured by the AFQT, to 22.1 as measured by the PAFQT; or the AFQT percentage is increased by about three-fourths. The increased frequency in categories IVB, IVC, and V is taken from the frequency in category IIIB (percentile scores 31 through 49).

Another estimate of the amount of compromise is available from the time when ACB-73 was first used (July 1973) to obtain AFQT categories. The percentage in category V increased from about 5.5 in June 1973 to about 9 in July (details are presented in appendix G). The proportional increase is about two-thirds. The proportional increase in categories IVB and IVC cannot be estimated directly. The increase from April 1973, the month preceding any use of ACB-73, to July 1973, when the use of AFQT 7/8 was suspended, is from about 13.5 percent to about 19.5 percent; the proportional increase was just under one-half.

A third estimate is that when ASVAB 5/6/7 replaced ACB-73, the increase in category V was from about 9 percent in December 1975 to about 16 percent in January 1976; the proportional increase was about three-quarters. Only the percentages for category V are cited because

the correct and original ASVAB 5/6/7 scales are in agreement only for category V.

The evidence is fairly consistent: to adjust for coaching, the percentage in AFQT categories IVB, IVC, and V should be increased by about two-thirds to three-quarters. In our adjustment for compromise, we increased the percentage of Army examinees in AFQT categories IVB, IVC, and V by 70 percent and we reduced the percentage in AFQT category IIIB accordingly. We chose 70 percent because it is about midway between two-thirds and three-quarters. We did not change the percentages in AFQT categories IVA, IIIA, II, or I.

We adjusted the distribution of AFQT scores from ACB-73 for the Army examinees tested during September and October 1975 (shown in appendix G) and computed an ASVAB 5/6/7 score scale (table 3).* The effects of coaching on ACB-73 inflated the original ASVAB 5/6/7 scale by an estimated 10 percentile scores at ASVAB 5/6/7 raw scores 27 and 28. At the bottom of the scale, below a raw score of 14, and at the middle of the scale, above a raw score of 41, the estimated effects of coaching were nil. Coaching on the AFQT scores from ACB-73 had a large effect on the inflated original scale.

Coaching, however, does not explain all the inflation of the original scale. In the last column of table 3, we show the correct scale, and as can be seen the scale adjusted for coaching on ACB-73 is still inflated. Thus, we continue our search for errors by turning to the testing procedures used for the Army examinees.

Faulty Testing Procedures

As we discussed in the Introduction to this chapter, the testing procedures were designed to minimize the burden on AFEES personnel and on the exmainees. To that end, operational AFQT scores from ACB-73 were used as the reference test for calibrating ASVAB 5/6/7. We found that coaching on the operational AFQT contributed up to an estimated 10 points to inflation of the original ASVAB 5/6/7 scale. Using the operational AFQT scores as the reference test could contribute to the inflation for two additional reasons:

- AFEES personnel probably tended to administer ASVAB 5/6/7 to those who had already taken ACB-73

* The rationale for restricting the Army sample to examinees tested in September and October 1975 is explained in the following subsection.

TABLE 3

EFFECTS OF COACHING IN THE ARMY SAMPLE
ON THE ASVAB 5/6/7 SCORE SCALE

ASVAB 5/6/7 raw score	Percentile score			
	Original[a]	Adjusted[b]	Difference	Correct
0-11				
12				
13	5			
14	6			
15	6	5	-1	
16	6	5	-1	
17	7	5	-2	2
18	8	6	-2	3
19	9	7	-2	4
20	10	8	-2	5
21	11	9	-2	6
22	13	9	-4	7
23	14	10	-4	9
24	17	12	-5	10
25	19	13	-6	11
26	22	14	-8	12
27	26	16	-10	13
28	28	18	-10	14
29	29	21	-8	15
30	31	23	-8	16
31	33	26	-7	17
32	34	29	-5	18
33	36	31	-5	19
34	37	33	-4	21
35	38	35	-3	23
36	40	37	-3	25
37	41	39	-2	27
38	43	41	-2	29
39	44	42	-2	31
40	46	44	-2	33
41	47	45	-2	35
42	48	48	0	38

[a]Scale based on original distributions of ACB-73 scores for examinees tested September and October 1975.
[b]ACB-73 scores adjusted for coaching on AFQT.

- Examinees are inclined to try harder on the operational test that determines their qualification for enlistment than on the experimental test.

Selecting Failures on the Operational Reference Test

To the extent that AFEES personnel administered ASVAB 5/6/7 only to those who had already taken ACB-73, the sample of Army examinees would be restricted in the upper end, above an AFQT score of 50, and in the low end, below an AFQT score of 21. Restriction at the upper end is less of a problem because the ASVAB 5/6/7 score scale at that end was established primarily by Navy and Air Force recruits. At the low end, however, the restriction would result in an inflated ASVAB 5/6/7 score scale. If applicants who fail to qualify for enlistment are systematically excluded from the calibration sample, then the distribution of reference test scores is biased too high. The reason is that measurement error, inherent in all aptitude tests, is primarily positive for those marginally qualified for enlistment; those with large negative errors of measurement tend to fail their enlistment tests. When those who are marginally qualified are given the experimental test, negative errors of measurement will occur on the experimental test, and the net effect is to lower the experimental scores relative to the reference test. To the extent that the testing procedures at the AFEES resulted in excluding failures on the reference test from the calibration sample, the ASVAB 5/6/7 score scale at the low end will be inflated.

The ASVAB 5/6/7 score distributions for the Army sample are shown in table 4 along with those for all male Army applicants in FY 1975. Comparison of the FY 1975 applicants and the total sample indicates that ASVAB 5/6/7 was given to a selected group of applicants. The percentage of the calibration sample with ACB-73 score in categories IIIA and above was only about one-half that of the FY 1975 applicants in these categories (about 20 percent versus 40 percent). This result suggests that applicants with AFQT-73 scores above 50 were systematically excluded from the calibration sample.

At the low end a similar result was found. The percentage of persons in categories IVC and V was less than for the total number tested in FY 1975. This result suggests that applicants who failed ACB-73 were systematically excluded from the calibration sample. The ACB-73 score distribution (disproportionately large number of examinees in category IIIB and disproportionately small number in categories IVC and V) indicates that the reference test scores for the calibration sample were not representative of the normal flow of Army applicants for enlistment.

The AFQT distributions for Army examinees grouped by date of testing with the ACB-73 are also shown in table 4. Group 1, which had scores of record dated before 1 September 1975, contained 556 cases, or 22 percent of the total Army sample. This group probably returned to the AFEES for further processing during the period of experimental

testing, and was then given ASVAB 5/6/7. Only 1 percent of this group had AFQT scores of record below the 10th percentile (category V). In contrast, about 7 percent of examinees with AFQT scores of record dated in September and October (groups 2 and 3) scored in category V.

TABLE 4

DISTRIBUTION OF ACB-73 SCORES FOR ARMY EXAMINEES IN CALIBRATION SAMPLE COMPARED TO FY 1975 ARMY APPLICANTS

AFQT category	FY 1975 Army applicants	Army calibration sample			
		Total[a]	Group 1[b]	Group 2[c]	Group 3[d]
I, II, IIIA	40.8	19	13	22	25
IIIB	27.4	49	70	40	43
IVA	10.3	14	11	15	14
IVB	5.7	5	3	7	5
IVC	7.5	7	2	8	7
V	8.3	6	1	8	6

[a]N = 2,512 Army male nonprior service applicants.
[b]N = 556, tested 1 January through 31 August 1975.
[c]N = 519, tested 1 through 15 September 1975.
[d]N = 632, tested 16 September through 31 October 1975.

We computed an ASVAB 5/6/7 score scale for the Army sample grouped by date of testing. The results for groups 1, 2, and 3 are presented in figure 5. (The cumulative frequency distributions for each group, including group 4, which has ACB-73 scores of record dated after October 1975, and the conversion tables are in appendix E.) The ASVAB score scale based on group 1 (tested with the ACB-73 before data collection began at the AFEES) is substantially higher in the low end (AFQT categories IV and V) than for groups 2 and 3, whose ACB-73 scores of record are dated later. The score scales for group 1 merge with those for groups 2 and 3 (tested in September and October 1975) above a percentile score of 30. In group 1 the ACB-73 and ASVAB 5/6/7 tests were likely administered under such different conditions that the two sets of scores are hardly comparable. Examinees in group 1 are best deleted from the calibration sample. Group 4 should be deleted because the ACB-73 scores of record are dated after most of the experimental testing was completed at the AFEES, and we do not know where the ACB-73 scores coded on the ASVAB 5/6/7 answer sheets came from.

Accordingly, we have deleted groups 1 and 4. Because in groups 2 and 3 both ACB-73 and ASVAB 5/6/7 were more likely to be administered on the same day, the data for these two groups should be more trustworthy.

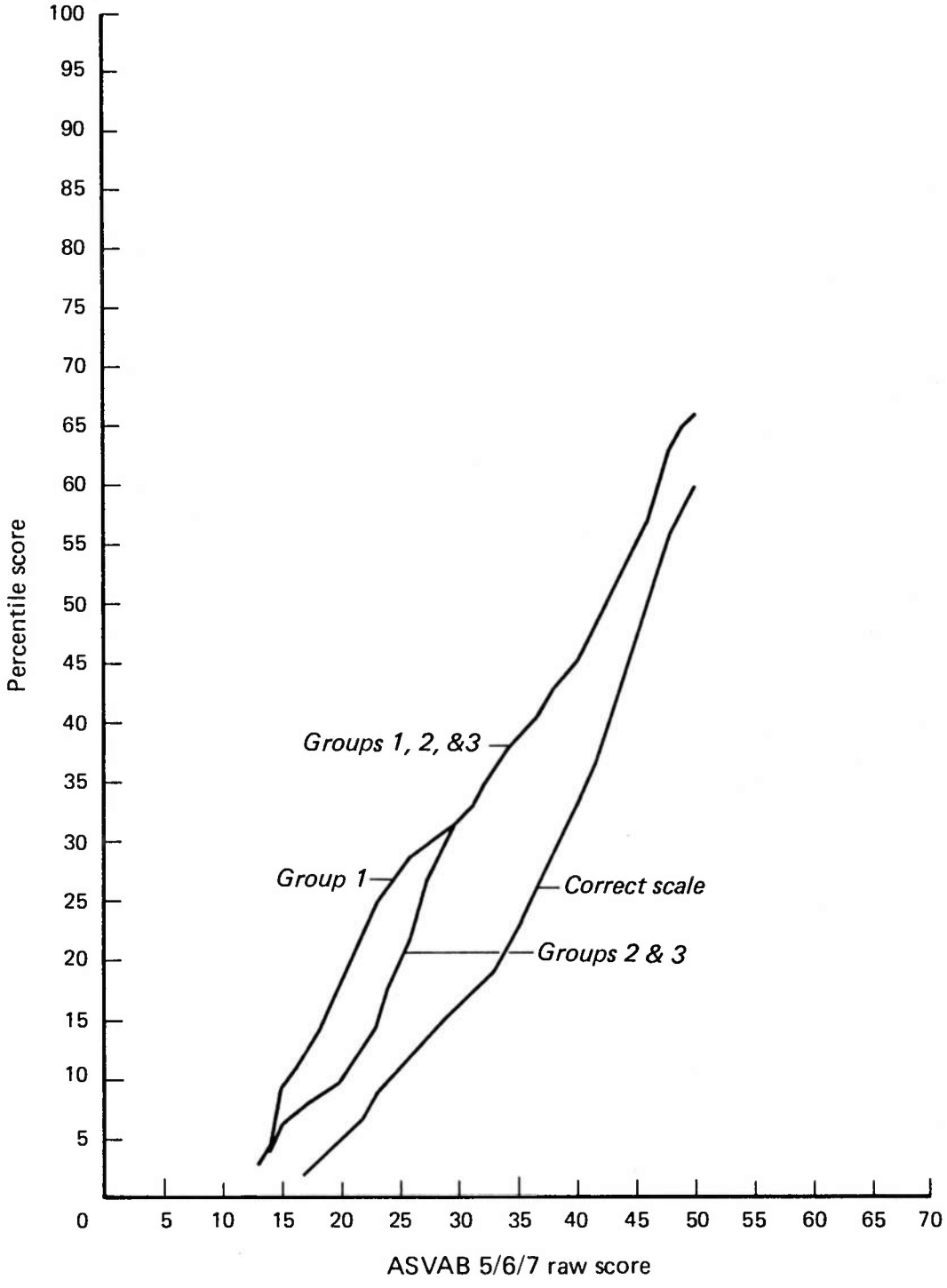

FIG. 5: CONVERSION OF ASVAB 5/6/7 RAW SCORE TO PERCENTILE SCORE FOR ARMY SAMPLE GROUPED BY DATE OF TESTING WITH ACB-73

The size of the Army sample for our purposes therefore was reduced to 1,151 cases (519 from group 2 and 632 from group 3). Also, we used the ACB-73 scores of record as the reference variable rather than the scores coded on ASVAB 5/6/7 answer sheets.

Examinees' Level of Effort When Taking ASVAB 5/6/7

In 1975 all of the Army examinees took ACB-73 because they wanted to qualify for enlistment and their level of effort should have been high. Their level of effort on the experimental ASVAB 5/6/7 probably was lower. Some reasons for this are:

- Many examinees knew they had already qualified for enlistment, and their scores on the experimental tests could not affect their qualification.
- There were obvious differences between the appearance of the operational and experimental test booklets. The operational booklets looked like official, professional instruments. The experimental booklets, in contrast, had items pasted on some of the pages; some sheets were a different color from the rest of booklet; some test booklets were stapled in such a way that some item numbers could not be read.
- Many examinees undoubtedly were told by their recruiters which tests counted for enlistment and which did not.
- Some applicants were retained for experimental testing, while others processing at the same time were released.
- Much of the experimental testing probably took place at odd hours, such as evenings, to fit into one day of processing.

The combined effect of these factors on the level of effort of Army examinees when taking the experimental tests could be substantial. The effect would be to lower the ASVAB 5/6/7 raw scores.

Estimated Effects on Inflation of the ASVAB 5/6/7 Scale.

In table 5 we show the adjustment to the ASVAB 5/6/7 scale arising from selection of Army examinees on the basis of their operational reference test scores. (Details are presented in appendix F.) The largest adjustment, 7 percentile score points, occurs at ASVAB 5/6/7 raw scores of 34 and 35, which convert to percentile scores of 21 and 23 in the correct scale. Above a raw score of 46, percentile score of 50, the adjustments are trivial and unreliable. We conclude that the combination of using the operational ACB-73 as the reference test plus the tendency to select examinees on the basis of their ACB-73 scores explains up to one-half the inflation of the original ASVAB 5/6/7 scale.

TABLE 5

ADJUSTMENT OF ASVAB 5/6/7 SCALE FOR PRIOR SELECTION OF EXAMINEES ON OPERATIONAL REFERENCE TEST

ASVAB 5/6/7 raw score	Percentile score[a]	Inflationary effect[a]
18	3	1
19	4	1
20	5	1
21	6	1
22	7	1
23	9	1
24	10	2
25	11	2
26	12	2
27	13	3
28	14	4
29	15	5
30	16	6
31	17	6
32	18	6
33	19	6
34	21	7
35	23	7
36	25	6
37	27	6
38	29	5
39	31	4
40	33	3
41	35	3
42	38	1
43	41	1
44	44	1
45	47	1
46	50	0

[a]On correct scale.
[b]Obtained from table F-3, appendix F.

Adusted ASVAB 5/6/7 Scale in Sample of Army Examinees

The inflationary effects of coaching and prior selection of examinees on the operational reference test were large. In table 6 we present score scales computed after adjusting the scores for different sources of errors. In figure 6 we graphed the amount of inflation accounted for by each effect. The total inflation was determined by subtracting the correct scale from the scale we computed on the full Army sample. A small residual of unexplained inflation remained below a raw score of 23, or in category V. A large amount of unexplained inflation, up to 9 percentile score points, remained above a raw score of 30; the largest residual occurred in the range of categories IVA and IIIB based on the correct scale.

SUMMARY

Our attempt to solve the problem of the inflated ASVAB 5/6/7 score scale produced some plausible answers. The first reason we found is that the reference tests for Navy and Air Force recruits most likely were not scored correctly. Then we started a painstaking examination of the data for Army examinees. We found three plausible reasons:

- Coaching on the reference test
- Selecting examinees on the basis of their reference test scores
- Lacking motivation when taking the experimental test.

The combined effect of these reasons was to explain almost all the inflation below a percentile score of 17 (raw score of 31). A large unexplained inflation persisted, however, between raw scores of 31 and 46.

We were unable to find computational or clerical-type errors that explained the inflation. We looked for arithmetic errors when computing the original scale on the sample of 1,600 cases. We also checked to see if bias occurred when the sample was reduced from over 5,000 cases to 1,600 or if inclusion of females in the calibration sample helped produce the error. We also examined the accuracy of the scoring keys. Because these simple and obvious possibilities did not work, our reanalysis of the Army examinees took less direct paths. Our explanations therefore involve an element of conjecture rather than being obvious and straightforward.

TABLE 6

SCALING ASVAB 5/6/7 IN THE SAMPLE
OF ARMY EXAMINEES

	Percentile score				
	A	B	C	D	
Raw score[a]	Full Army sample[b]	Tested Sep/Oct[c]	Coaching[d]	Prior selection[e]	Correct[f]
15	6	6	5	5	
16	7	6	5	5	
17	7	7	5	5	2
18	8	8	6	6	3
19	10	9	7	7	4
20	11	10	8	7	5
21	13	11	9	8	6
22	15	13	9	8	7
23	17	14	10	9	9
24	19	17	12	10	10
25	21	19	13	11	11
26	24	22	14	12	12
27	27	26	16	13	13
28	29	28	18	14	14
29	30	29	21	16	15
30	32	31	23	17	16
31	33	33	26	20	17
32	35	34	29	23	18
33	36	36	31	25	19
34	37	37	33	26	21
35	38	38	35	28	23
36	40	40	37	31	25
37	41	41	39	33	27
38	42	43	41	36	29
39	43	44	42	38	31
40	45	46	44	41	33
41	46	47	45	43	35
42	47	48	48	47	38
43	48	51	51	50	41
44	49	53	53	52	44
45	50	55	55	54	47
46	54	58	--	--	--

[a]Scale extends only to raw score of 46 because data unreliable above that point.
[b]Scale based on all Army examinees, N = 2,512.
[c]Based on examinees tested with ACB-73 from September through October 1975, N = 1,151.
[d]Column C scale adjusted for coaching on ACB-73.
[e]Column D scale adjusted for prior selection of examinees on ACB-73.
[f]Correct scale, adopted in 1980.

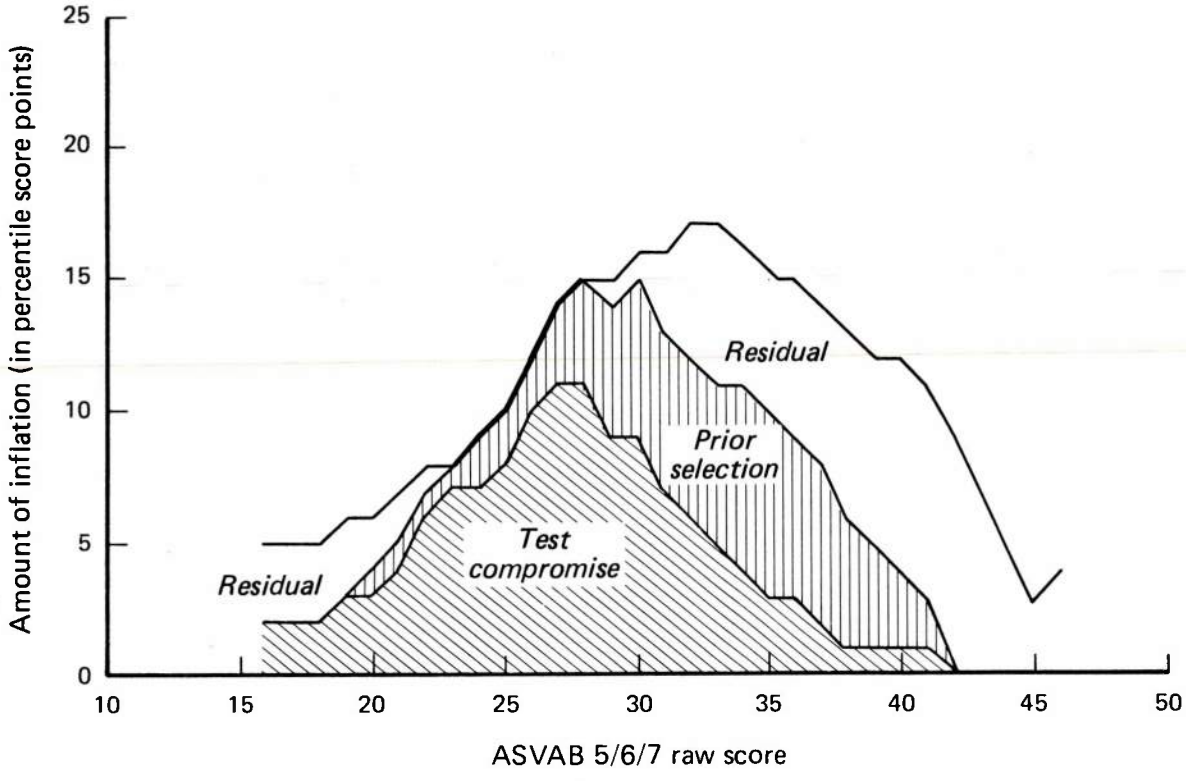

FIG. 6: AMOUNT OF INFLATION IN ASVAB 5/6/7 SCORE SCALE EXPLAINED BY ADJUSTMENTS TO THE ARMY SAMPLE

CHAPTER 3

REPRODUCING THE CORRECT SCALE IN THE ORIGINAL CALIBRATION SAMPLE

In our analysis of the original calibration sample we made two adjustments to the reference test scores: we rescored the ASVAB 2 reference test for Navy and Air Force recruits, and we estimated the amount of test compromise on ACB-73 for the Army examinees. A third adjustment was to the scale itself: we adjusted the percentile scores for prior selection of Army examinees on the basis of their operational reference test scores. Each of these adjustments is independent of the others. Therefore, the effects on the inflated ASVAB 5/6/7 scale are cumulative. We applied each adjustment in turn to the combined samples. The Army sample was limited to examinees tested with ACB-73 during September and October 1975.

COMPARISON OF SCORE SCALES

The resulting score scales are shown in table 7. (The cumulative frequency distributions used to compute the score scales are in appendix C.) In our attempt to reproduce the original scale, we found that the original scale was too difficult below an ASVAB 5/6/7 raw score of 36; our recomputed scale, shown in column B, was up to 11 percentile score points higher, at a raw score of 29, than the original scale. Above a raw score of 37, the original scale tended to be more inflated than the recomputed scale; the maximum difference was 5 percentile score points at a raw score of 41. Because we could not reproduce the original scale, and because the recomputed scale is consistent with the data, we must explain the differences between the correct and recomputed scales (columns F and B in table 7).

The effect of rescoring ASVAB 2 for the Navy and Air Force recruits was to explain most of the inflation above a raw score of 50. Rescoring explained some of the inflation as low as a percentile score of about 20. The scale based on rescoring ASVAB 2 is shown in column C. The amount of inflation explained by the rescored ASVAB 2 is diagrammed in figure 7. A relatively small residual of unexplained inflation, 2 percentile score points, remains above a raw score of 58. At the upper end, the scale is unreliable because the scaled scores are based on only a relatively few cases.

Column D of table 7 shows the effects of adjusting the ACB-73 score distribution for test compromise. The scale in column D also includes the effects of rescoring ASVAB 2. As shown in figure 7, the effects of test compromise are most pronounced between raw scores of about 20 and 36. In this interval, test compromise explains an inflation effect of up to 6 percentile score points.

TABLE 7

ADJUSTED ASVAB 5/6/7 SCALES IN 1975 CALIBRATION SAMPLE

	Percentile score					
	A	B	C	D	E	F
ASVAB 5/6/7 raw score	Original[a]	Recomputed[b]	Rescored ASVAB 2[c]	Compromised[d]	Prior selection[e]	Correct[f]
0-12	1					
13	2	3				
14	2	5				
15	2	6				1
16	3	7				1
17	3	8	7	6	5	2
18	3	9	8	7	6	3
19	4	10	9	8	7	4
20	6	12	10	9	8	5
21	7	14	11	9	8	6
22	8	16	12	10	9	7
23	9	18	14	11	10	9
24	11	20	18	13	11	10
25	13	23	20	14	12	11
26	17	26	21	15	13	12
27	18	28	23	17	14	13
28	21	30	25	19	15	14
29	21	32	27	21	16	15
30	25	33	28	23	17	16
31	30[g]	35	30	25	19	17
32	33	36	31	27	21	18
33	36	38	33	30	23	19
34	38	40	35	32	25	21
35	42	41	37	34	27	23
36	43	43	38	36	30	25
37	45	45	39	38	32	27
38	48	47	40	39	34	29
39	49	48	42	41	37	31
40	54	50	44	42	39	33
41	58	53	46	45	43	35
42	60	56	48	48	47	38
43	62	59	49	49	48	41
44	64	62	51	51	50	44
45	65	64	53	53	52	47
46	67	66	55	55	55	50
47	69	68	57	57	57	53

TABLE 7 (Cont'd)

	Percentile score					
	A	B	C	D	E	F
ASVAB 5/6/7 raw score	Original[a]	Recomputed[b]	Rescored ASVAB 2[c]	Compromised[d]	Prior selection[e]	Correct[f]
48	71	70	59	59	59	56
49	74	72	60	60	60	58
50	75	75	62	62	62	60
51	77	77	64	64	64	62
52	79	78	66	66	66	65
53	80	80	68	68	68	67
54	82	82	70	70	70	70
55	84	84	73	73	73	72
56	85	85	76	76	76	75
57	87	87	79	79	79	77
58	88	89	81	81	81	80
59	89	90	84	84	84	82
60	92	91	86	86	86	84
61	93	92	88	88	88	86
62	94	93	89	89	89	87
63	95	93	91	91	91	89
64	96	94	92	92	92	91
65	97	95	94	94	94	93
66	98	96	95	95	95	95
67	98	97	96	96	96	97
68	99	98	98	98	98	98
69-70	99	99	99	99	99	99

[a]Original scale implemented 1 January 1976.
[b]Scale computed on a full sample of 4,588 males using equipercentile equating technique.
[c]ASVAB 2 reference tests for Navy and Air Force recruits rescored.
[d]ACB-73 reference test scores for Army examinees adjusted for test compromise.
[e]Scale adjusted for prior selection of Army examinees on basis of their reference test scores.
[f]Correct scale adopted by DoD in 1980.
[g]Changed to 31 in February 1976.

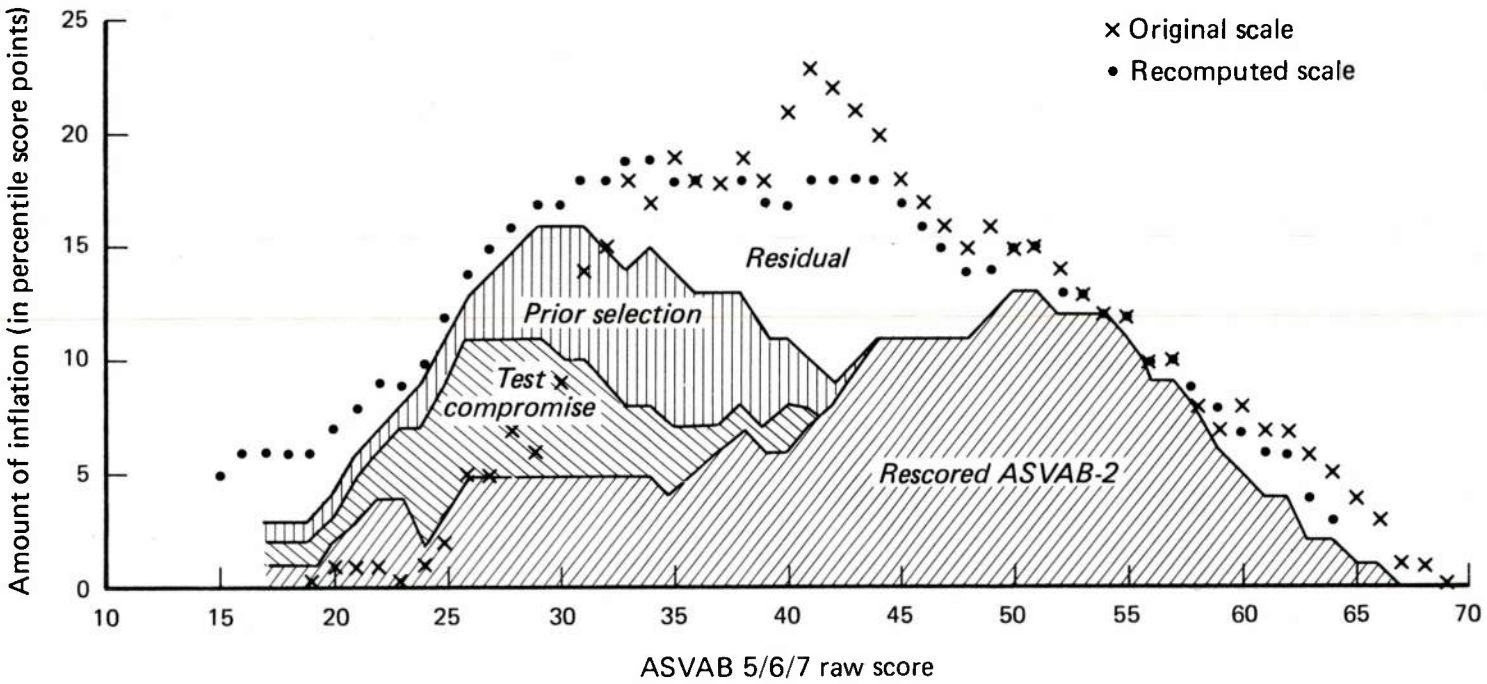

FIG. 7: AMOUNT OF INFLATION EXPLAINED BY ADJUSTMENT TO THE FULL ASVAB 5/6/7 CALIBRATION SAMPLE

The third effect, prior selection on the operational reference test, had its greatest effect on the raw score ranges 26 through about 40. As shown in columnn E of table 7 and in figure 7, prior selection explains an inflation effect of up to 7 percentile score points. Neither prior selection nor test compromise for the Army sample has a consistent effect above a raw score of 42. Therefore, we did not compute the effects of these adjustments above that point.

The cumulative effect of the three adjustments still leaves a residual of unexplained inflation. At both extremes, below a raw score of about 23 and above a raw score of about 59, the residual may be a function of unreliable conversions because of insufficient data. In the midrange, raw scores 30 to 52, the residual is large, reaching a maximum at a raw score of 42, where the difference between the correct scale and the final adjustment, shown in column E, is 9 percentile score points. Another way of looking at the residual inflation is that it corresponds to between 2 and 3 raw score points. Because ASVAB 5/6/7 has only about 55 useful points of discrimination, between raw scores 15 and 70, each raw score in the midrange corresponds to 3 percentile score points.

As we found in chapter 2, when we analyzed the samples of Navy and Air Force recruits and Army examinees separately, the three adjustments explain most of inflation from raw scores 23 to 30 and 53 to 59, while leaving a relatively large residual in between. In the remainder of this chapter, we discuss three possible errors in the score scales that could explain the residual inflation:

- Difficulty of correct scale--the correct scale may be too difficult in the raw score range 31 to 52. If the correct scale were too difficult, then raw scores in this range should convert to higher percentile scores.

- Inflation of reference test scores--the scales for ACB-73 and ASVAB 2 may themselves be inflated in this range. If these reference tests had inflated scales, then the ASVAB 5/6/7 scale would, to that extent, also be inflated.

- Existence of other factors--other factors may have been operating over the years to change the meaning of the score scale. The residual inflation could have arisen from changes in the content of AFQT, changes in the samples used to calibrate replacement forms of the AFQT, or changes in the mobilization population since World War II.

Is the Correct Score Too Difficult?

The available evidence indicates that the correct scale is, in fact, accurate:

- In May 1976, the Navy and Air Force calibrated ASVAB 5/6/7 to AFQT 7/8 on samples of recruits. The results (presented in appendix A) were virtually identical to the correct scale in categories I, II, III, and IVA. Insufficient data were available for reliable conversions below a percentile score of 20.

- In 1979, three independent studies, based on samples of service applicants, Marine Corps recruits, and high school students, produced essentially the same scale. Results from these studies were used to establish the correct scale.

- In October 1980, the transition from ASVAB 5/6/7, correctly scaled, to ASVAB 8/9/10 showed little change in the score distributions. The score distributions of applicants for each service before and after ASVAB 8/9/10 was introduced are shown in table 8. The calibration of ASVAB 8/9/10 is described in appendix H. Given the care with which ASVAB 8/9/10 was calibrated to AFQT 7A, the scale for ASVAB 8/9/10 is accurate, and, by inference, so is the correct scale for ASVAB 5/6/7.

The evidence supporting the accuracy of the correct scale for ASVAB 5/6/7 is strong, and we conclude that the correct scale is, in general, accurate.

Are the Scales for ACB-73 and ASVAB 2 Inflated?

If the scales for the two reference tests, ACB-73 and ASVAB 2, are themselves inflated, then the scale ASVAB 5/6/7 will also be inflated. In appendix D we describe the calibration of ACB-73, and we present data on the accuracy of the scales for forms 2 and 3 of the ASVAB. Forms 2 and 3 were parallel forms, developed and calibrated simultaneously. Form 3 was used by the Marine Corps and Air Force to select and classify recruits prior to introduction of forms 5, 6, and 7. Form 2 was used in the high school testing program. Because forms 2 and 3 of the ASVAB are parallel, they have the same score scale, and the accuracy of one form supports the other. The results presented in appendix D indicate that, in general, the scales for ACB-73 and ASVAB 2/3 are accurate. There are some indications, however, that the scales may be somewhat inflated in AFQT category IVA (percentile scores 21 through 30).

In table 9 we summarize the experience of the Army and Marine Corps when ACB-73 and ASVAB 3, respectively, were introduced by the services for selecting and classifying recruits.

TABLE 8

SCORE DISTRIBUTIONS OF MALE APPLICANTS WHEN ASVAB 8/9/10 WAS INTRODUCED

AFQT category	Percent in AFQT category					
	ASVAB 5/6/7			ASVAB 8/9/10[a]		
	July	Aug	Sept	Oct	Nov	Dec
Army						
I & II	13.5	15.4	17.2	18.8	18.8	17.3
IIIA	9.7	9.9	10.7	11.0	10.9	10.6
IIIB	15.2	15.3	15.0	17.5	17.6	16.5
IVA	14.9	14.6	12.8	16.1	13.6	13.2
IVB & C	32.6	31.7	31.9	22.4	24.7	26.8
V	14.1	13.1	12.4	14.2	14.4	15.6
Navy						
I & II	34.0	34.6	32.1	32.7	33.0	32.7
IIIA	18.9	18.4	18.6	18.0	18.0	17.5
IIIB	20.0	21.1	20.9	21.0	21.0	20.5
IVA	11.0	11.4	11.6	12.1	12.0	12.6
IVB & C	13.2	12.1	14.0	12.1	12.4	12.4
V	2.9	2.4	2.8	4.1	3.6	4.3
Air Force						
I & II	30.8	31.8	30.5	36.0	36.0	34.0
IIIA	18.7	19.2	19.5	20.9	20.2	19.3
IIIB	21.8	21.6	22.7	21.0	21.1	22.1
IVA	11.9	11.8	12.6	10.2	10.1	11.5
IVB & C	13.2	12.6	12.0	9.1	10.1	10.5
V	3.6	3.0	2.7	2.8	2.5	2.6
Marine Corps						
I & II	21.7	21.3	20.3	23.5	23.9	23.8
IIIA	15.2	15.5	15.6	15.3	16.3	15.2
IIIB	20.8	21.4	20.6	23.3	22.1	22.6
IVA	15.4	14.6	15.0	14.3	14.1	14.1
IVB & C	20.7	21.3	22.5	17.0	16.7	18.3
V	6.2	5.9	6.0	6.6	6.9	6.0

[a]ASVAB 8/9/10 implemented on 1 October 1980.

TABLE 9

AFQT SCORES OF ARMY AND MARINE CORPS APPLICANTS
WHEN AFQT 7/8 WAS REPLACED

AFQT category	Percent in category			
	Army applicants		Marine Corps applicants	
	AFQT 7/8[a]	ACB-73[b]	AFQT 7/8[c]	ASVAB 3[d]
V	5	9	6	10
IVB & IVC	18	19	17	20
IVA	14	11	16	17
IIIB	21	24	21	25
IIIA	17	15	18	13
I & II	24	21	22	16

[a]June 1973.
[b]July 1973.
[c]June 1974.
[d]July 1974.

For both batteries, the percentage of applicants placed in categories IVB, IVC, and V increased immediately after their introduction. This increase is expected because AFQT 7/8 had been used since 1960, and its content was readily available to recruiters. The percentage in category IVA for ACB-73 declined in July 1973 compared to the preceding months, while the percentage for ASVAB 3 remained relatively constant. It is possible that the effects of scale inflation and test compromise tended to balance each other in category IVA.

In category IIIB the percentage for both tests increased. Based on our earlier analysis of the compromise in the Army sample, where we compared AFQT and PAFQT scores, we would expect the percentage in category IIIB to decline when a compromised test is replaced by a new test. In addition, if the scales for ACB-73 and ASVAB 3 were inflated in category IIIB relative to AFQT 7/8, again we would expect the percentage to decrease. Instead of a decrease, there was an increase of about 4 percentage points for each service (from about 20 to 24 percent for the Army and 21 to 25 percent for the Marine Corps).

In categories IIIA and above, both sets of percentages declined. These declines may be a function of test compromise in the above average range. In chapter 4 we discuss this possibility at greater length.

The percentages in the upper categories for the Marine Corps applicants tended to resume their former levels after ASVAB 3 had been in operation for 3 months. For Army applicants, the percentage in category IIIA tended to reach its former level (about 17 percent of all applicants). But, in categories I and II, the percentage remained at 21 for July through September, in contrast to 28 percent during April and May and 24 percent in June. These data support that there was some inflation of the ACB-73 and ASVAB 3 scales in the category IVA range, but not in the category IIIB range.

Some of the residual inflation between ASVAB 5/6/7 raw scores of 30 and 39 could be explained by inflation of the ACB-73 and ASVAB 3 scales. Most of the residual inflation, as shown earlier in figure 7, is above a raw score of 39, and our conclusion is that the residual above a raw score of 39 is probably not explained by inflation of the ACB-73 and ASVAB 3 scales.

Do Other Factors Explain the Residual Inflation?

The ASVAB score scale has a history dating from inception of the first AFQT in 1950. The scale for the first AFQT was based on the score distribution of all men serving during WWII. During the two decades between WWII and the calibration of ACB-73 and ASVAB 2/3, many changes took place in AFQT content, population, and calibration procedures. We discuss each of these changes in turn.

Changes in Test Content

The first form of the AFQT and the tests used during WWII had three types of items: word knowledge, arithmetic reasoning, and space perception. AFQT forms used from 1953 through the early 1970s, forms 3 through 8, had four types of items--the same three as the first form, plus knowledge of tool functions. The items on knowledge of tools were dropped from the AFQT when ACB-73 and ASVAB 2/3 were introduced. All AFQTs used after AFQT 7/8 had only three types of items.

The presence or absence of the items on knowledge of tools could have an effect on the score scale. Because the content of the AFQT was different between AFQT 7/8 and subsequent versions, the score scale could have been changed when ACB-73 and ASVAB 2/3 were calibrated. When the reference and new tests have different content, then the scale is not uniquely determined; rather the scale depends on the characteristics of the calibration sample. The scales for ACB-73 and ASVAB 2/3 may be distorted to some degree depending on the distribution of scores for the tool knowledge items relative to the distributions for the other three types of items that are common to the forms (word knowledge, arithmetic

reasoning, and space perception). The difference in test content could explain part of the residual inflation.

Changes in Population

When the scale for AFQT 7/8 was developed in 1959, the calibration sample consisted of registrants for the draft and relatively few minorities. The number of minorities in the sample was probably between 10 and 15 percent. During WWII when the reference population was created, the number of minorities was less than 10 percent. In 1979 when the correct scale for ASVAB 5/6/7 was developed, the calibration sample included about 35 percent blacks plus about another 10 percent of other ethnic minorities. Although racial composition of the calibration samples has been shown to have little effect on the ASVAB score scale, the effects could contribute to a shift in the score scale [2].

In addition to racial composition, the population has also changed in level of education since WWII. Level and type of education could affect the score scale.

Changes in the population could interact with changes in test content. If content of the reference test were the same as content of the new tests, then changes in the population would probably have little effect on the scale. But with the deletion of tool knowledge items from AFQT, the interaction effect with population changes could explain part of the residual inflation.

Changes in Test Calibration Procedures

Perhaps the most important change was in the procedures used to calibrate the new tests. The calibration of ACB-73, ASVAB 2/3, and ASVAB 5/6/7 is suspect because shortcuts were used to calibrate them. In all three cases, the operational AFQT scores were used as the reference variable. The operational AFQT scores may have been inflated by test compromise, and some examinees may have tried harder on the operational test than on the experimental tests. The effect from both sources is to inflate the scale of the new test. The residual inflation could be explained in part by inflated scales that result from using shortcuts to calibrate the tests.

A shortcut in the calibration procedure was also used when the correct ASVAB 5/6/7 scale was developed. When these data were collected in June and July 1979, the operational ASVAB 5/6/7 scores were used as the "new" test, and AFQT 7A was administered as the "experimental" reference test. To the extent that applicants in the calibration sample tried harder on the operational ASVAB than on the AFQT 7A reference test, the correct scale is too difficult. Therefore, even though the evidence is strong that the correct scale is, in fact, accurate, the procedures could have led to a scale that is too difficult. We cannot rule out the possibility that the correct scale is too difficult, which could explain part of the residual inflation.

The residual inflation could also be explained in part by imprecise adjustments to the Army sample. We estimated that test compromise changed the score distributions in categories V, IVB, and IVC such that the observed distribution should be increased by 70 percent. Also, we used the 1980 ASVAB 8 calibration sample to estimate the combined effects of using the operational AFQT as the reference test and of tending to select failures on the AFQT. We suspect that the latter estimate is too conservative, and we have no good independent check of the amount of compromise in the 1975 calibration sample. There simply are no conclusive explanations for the residual inflation.

SUMMARY

Our attempt to reproduce the correct scale in the full original ASVAB 5/6/7 calibration sample for the most part succeeded. We found that three reasons explained most of the inflation.

A residual inflation of up to 9 percentile score points or up to about 3 raw score points, however, remains unaccounted for. We considered three possible explanations for the residual inflation: inflated scales for ACB-73 and ASVAB 2/3, the original reference tests for calibrating ASVAB 5/6/7; changes in content of the AFQT and in the mobilization population; and shortcuts in the calibration procedures. Each of these could have some effect on the ASVAB 5/6/7 scale that results in the residual difference between the correct scale and the original scale.

We are still uncertain about the meaning of the ASVAB 5/6/7 scale. We cannot conclude definitely that the correct scale, in fact, matches the WWII reference population. We came close to reproducing the correct scale, but the residual inflation raises doubts about the meaning of all the score scales used, except AFQT 7/8.

In the next chapter we discuss problems in interpreting the score distributions, including those for AFQT 7/8, obtained during the AVF era.

CHAPTER 4

WHY WAS INFLATION MASKED AT THE BELOW-AVERAGE RANGE OF SCORES?

INTRODUCTION

When ASVAB 5/6/7 was introduced in January 1976, the ASVAB Working Group compared the distribution of AFQT scores for recruits obtained on the new test with scores of recruits immediately prior to its introduction. Based on the data, and assumptions about the accuracy of the scores, they concluded that the ASVAB 5/6/7 scores may be inflated. They initiated studies to check the accuracy of the scores, and by May 1976, they had verified that the scale indeed was inflated in the above-average range. They also concluded that the scores in the low range were sufficiently accurate and that no adjustment to the scale was required.

In this chapter we state three assumptions required to evaluate the accuracy of scores on a new test by comparing score distributions on the new and the previous tests.* We then examine how tenable the assumptions were in 1975 and how failure to meet the assumptions could have led to false conclusions about the accuracy of the ASVAB 5/6/7 scale. We conclude the chapter by discussing the accuracy of all test score distributions during the first years of the AVF era (1973 through 1975).

ASSUMPTIONS FOR COMPARING SCORE DISTRIBUTIONS

Three assumptions are required for using score distributions before and after introduction of a new version as the basis for evaluating the accuracy of the new score scale:

- Scores for the new test are accurate; that is, the scale is calibrated correctly to the reference population and the scores are computed correctly.

---

* The long-term check on the accuracy of the score scale is changes in failure rates, especially during skill training, that may reflect an inadvertent change in standards. If the scale has shifted, then the standards have also shifted accordingly, and the effects should be apparent in failure rates. During the period when ASVAB 5/6/7 was operational, the Army raised aptitude composite prerequisites in more than 50 skill training courses. The Army experience was that failure rates in many skills were becoming excessive, and the solution was to raise the standards for assignment to those skills. The inflated ASVAB scale of course inadvertantly lowered standards, and the Army merely restored standards for these skills to former levels.

- Ability of applicants does not change when the new test is introduced.

- Scores for the old tests are accurate; that is, the scale for the old test is correct, the amount of test compromise is negligible or known, and the scores are computed correctly.

During the draft era, score scales for new forms of the AFQT could easily be verified by comparing distributions of test scores immediately before and after the new test was introduced. Distributions were obtained for young men registering for induction. Test compromise was not a problem among the registrants, and the scores of the registrants ordinarily showed little fluctuation from month to month. The historical evidence is that the transition for each new version of the AFQT was smooth; the accuracy of the AFQT score scale was not a subject of controversy during the draft era.

However, in 1976 when ASVAB 5/6/7 was introduced, the situation was different. All examinees were applicants for enlistment. As we found for the Army sample, many of them were coached on the test, which means that the score distributions on the previous test were inflated. Also as we found in chapter 2, the calibration of some of the tests used in 1975 may be in doubt (ACB-73 and ASVAB 3). In retrospect, we now know that the baseline data for comparing the ASVAB 5/6/7 scale is not as solid as we would like for purposes of evaluating the accuracy of the ASVAB 5/6/7 scale. We examine each of these assumptions to see how they affected accuracy of the scores and consequently any misleading interpretation of the score distributions in 1975 and 1976.

EXPECTED EFFECTS OF INFLATED ASVAB 5/6/7 SCALE ON SCORE DISTRIBUTIONS

Before examining shifts in the 1975 and 1976 observed distributions of scores, we need to make a preliminary examination of how the inflated scale affected the actual score distributions for ASVAB 5/6/7 and what the score distributions would have been if the correct scale had been used. By comparing these distributions to those for the tests used in 1975, we can gain further insight into the accuracy of the scores and perhaps of why the ASVAB Working Group did not change the bottom end of the ASVAB 5/6/7 scale in September 1976.

The effects of the inflated scale are not uniform throughout the score range; the inflated scale affected some AFQT categories more than others. The key element that affects the amount of inflation of the scale is the number of test items or raw score points in each AFQT category. There is a strong relationship between the percentage of examinees in a category and the number of points. A secondary element that affects the amount of inflation is the location of the AFQT category in the scale. As a rule, more examinees score in the middle of the range than at the high or low ends. These two elements, plus the

ability of the examinees relative to the reference population, explain the observed score distributions.

The percentages of examinees placed in the AFQT categories must sum to 100. If one category has an excess of cases, then another must be deficient. For example, the ASVAB 5/6/7 score scale was known to be inflated at the upper end almost immediately after introduction. This meant that compared to the correct scale, an excessive number of applicants were placed in AFQT categories I and II. These people had to come from lower categories, which then must have a deficiency of cases. If the inflation were uniform, then examinees would be shifted upward out of categories IV and V and into categories I, II, and III.

Recall that the problem with the score scales stemmed from the observation that the top of the scale was inflated because there were too many applicants in categories I and II, and subsequently that the low end was inflated, leaving too few applicants in category IV. The observed score distributions should have shown a drop in the percent of cases in category IV immediately after ASVAB 5/6/7 was introduced. As discussed in the following paragraphs, such a shift was not found in the observed score distributions. Later questions arose about whether the correct scale was, in fact, accurate or whether the correct scale was too difficult at the low end.

DISTRIBUTION OF ARMY APPLICANTS ON THE ASVAB 5/6/7

The raw score boundaries of AFQT categories for the three ASVAB 5/6/7 score scales are shown in table 10. The percentages of Army male-nonprior-service applicants tested in CY 1976 placed in each AFQT category by the three scales are also shown in table 10. The percentages are based on only the CY 1976 applicants with ASVAB subtest scores in the automated data files, and, hence, may not agree exactly with other figures for the same period. Because cumulative distributions start at the low end of the scale, the low scores are listed first.

The original and correct scales agree exactly in category V in that they include the same raw score interval (0 to 23), but for some reason the operational scale (adopted in August 1976 and used through September 1980) removed one raw score point from category V. The biggest difference between the correct and other scales occurs in category IVC, percentile scores 10 through 15. The original scale contained only two items, whereas the correct scale contains six. The differences in percentage in Army applicants is in the same 1 to 3 ratio, 5.7 versus 18.2. Category IVB has a ratio of 1 to 2 (two items versus four and 6.2 percent versus 12.4) for the original and correct scales.

In categories IVA, IIIB, and IIIA the two scales have about the same number of raw score points. The percentage of applicants in category IIIB differs (21.8 for the original versus 14.7 for the correct), however, because the raw score interval for the original scale (32 through 39) is closer to the mean of Army applicants than is the raw

TABLE 10

PERCENT OF ARMY MALE APPLICANTS IN EACH AFQT CATEGORY USING ALTERNATIVE SCORE SCALES

AFQT category	Percentile score interval	Raw score interval	Score[a] scale	Number of raw score points	CY[b] 1976
V	1-9	0-23	Original	24	14.9
		0-22	Operational	23	12.4
		0-23	Correct	24	14.9
IVC	10-15	24-25	Original	2	5.7
		23-25	Operational	3	8.2
		24-29	Correct	6	18.2
IVB	16-20	26-27	Original	2	6.2
		26-27	Operational	2	6.2
		30-33	Correct	4	12.4
IVA	21-30	28-31	Original	4	12.6
		28-30	Operational	3	9.3
		34-38	Correct	5	13.4
IIIB	31-49	32-39	Original	8	21.8
		31-42	Operational	12	31.7
		39-45	Correct	7	14.7
IIIA	50-64	40-44	Original	5	10.6
		43-51	Operational	9	15.6
		46-51	Correct	6	9.8
II	65-92	45-60	Original	16	22.6
		52-63	Operational	12	13.6
		52-64	Correct	13	14.3
I	93-99	61-70	Original	10	5.6
		64-70	Operational	7	3.0
		65-70	Correct	6	2.3

[a]Original = original scale, used January to August 1976; Operational = operational scale, used September 1976 to September 1980; Correct = correct scale, adopted by DoD in 1980.
[b]Army applicants tested January to December 1976, N = 93,705 (only examinees with ASVAB 5/6/7 raw scores included).

score interval for the correct scale (39 through 45). In categories I and II the original scales contain many more items than does the correct scale (26 versus 19). The operational scale, used from September 1976 until October 1980, agreed with the correct scale in categories I and II, but it moved the excess items and examinees into categories IIIA and IIIB. In addition, raw score 31 was moved from category IVA into IIIB. The effect was that the operational scale placed about half of all Army applicants in category III.

Based on the number of raw score points in each category for the original scale, the following changes should have occurred in the observed score distributions of Army applicants between December 1975 and January 1976:

Category	Change
V	None
IVC	Decrease to a third of December 1975 percentage
IVB	Decrease to a half of December 1975
IVA	Little change
IIIB	Little change
IIIA	Little change
I and II	Increase by half

Based on the number of items or raw scores in each AFQT category we would expect the percentages in the lower end (percentile scores 10 through 20) to show a sudden decrease and the percentage at the upper end (percentile scores 65 through 99) to increase. These expectations should hold if the scores on the tests used in 1975 were accurate and if the ability of the examinees did not change between the end of 1975 and the beginning of 1976. We now turn to the observed distributions of scores during the transition to ASVAB 5/6/7 to see how they did, in fact, shift and how they would have shifted if the correct scale had been introduced.

ASVAB 5/6/7 SCORE DISTRIBUTIONS IN EARLY CY 1976

In early CY 1976 the ASVAB Working Group and personnel managers did not have the correct scale available for evaluating the ASVAB 5/6/7 distributions of scores. Neither did they have reason to suspect the accuracy of the scales for ACB-73 and ASVAB 3; nor did they know the extent to which the AFQT scores in late CY 1975 were inflated by test compromise. The main basis for evaluating the accuracy of the ASVAB 5/6/7 scale was the comparison of the observed score distributions in early CY 1976 with those in late CY 1975. The AFQT distributions for Army and Marine Corps applicants are shown in figure 8.

The changes in AFQT scores were similar for both services. The most startling change observed at that time was the sudden increase of

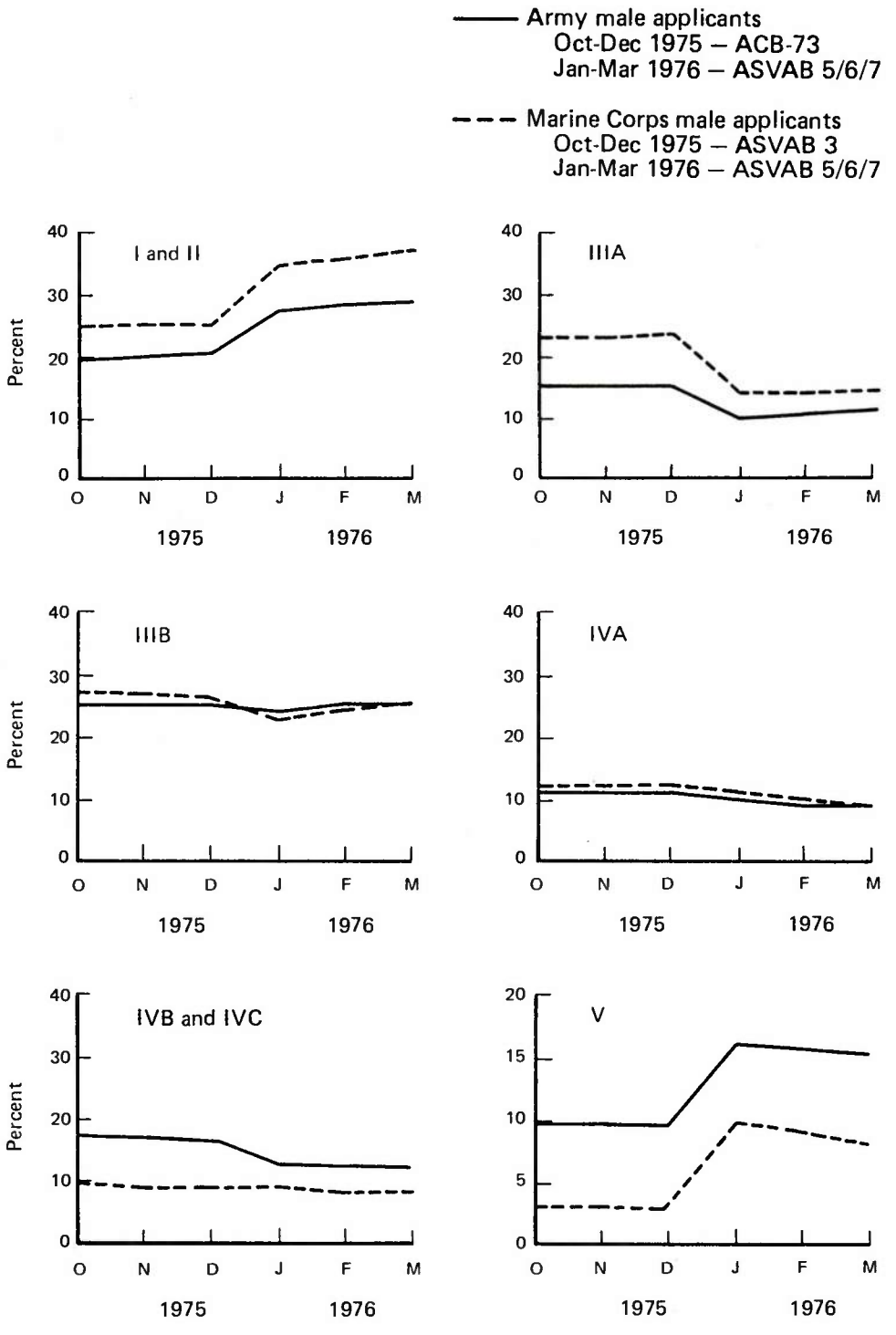

FIG. 8: OBSERVED SCORE DISTRIBUTIONS OF ARMY AND MARINE CORPS APPLICANTS, ORIGINAL ASVAB 5/6/7 SCORE SCALE

examinees in categories I and II. The immediate suspicion of the ASVAB Working Group was that the ASVAB 5/6/7 scale was inflated at the top. The Navy and Air Force conducted calibration studies in April and May 1976 (described in appendix A) and found that indeed the scale was inflated at the upper end. Given that the observed percentages in categories IIIA and IIIB declined in January 1976, a reasonable inference is that the increase in categories I and II came out of categories IIIA and IIIB.

The fix to the scale adopted in September 1976 made just such a change--the percentage of examinees placed in categories I and II was decreased, and the percentage in categories IIIA and IIIB increased. These changes to category III and the lack of changes in categories IV and V were not unduly suspicious. They could be explained by test compromise and perhaps some unreliability in the score scale.* In any event, no changes, except for minor adjustments, were made in 1976 to the ASVAB 5/6/7 scale in categories IV and V.

The ASVAB 5/6/7 distributions based on the correct ASVAB 5/6/7 scale for Army and Marine Corps applicants are shown in figure 9. If the correct scale had been used in January 1976, the percentage of applicants in categories IVB and IVC would have shown a large increase, double for the Army and almost triple for the Marine Corps. Category V would not have differed from the observed distribution because the original and correct scales were identical in category V. Category IVA would have shown some increase. Categories I, II, and III would have shown large decreases for both services.

If the correct ASVAB 5/6/7 scale had been introduced in January 1976, the sharp changes from the preceding month would have received much more attention from DoD personnel managers and the ASVAB Working Group. Because changes of such magnitude in a 1-month period are unexpected they would have led to an intense examination of the reason for the large changes. We explain the large shifts for the correct scale. We examine the implications of the three assumptions mentioned earlier. Each is discussed in turn.

DID THE ABILITY OF APPLICANTS CHANGE WHEN THE NEW TEST WAS INTRODUCED?

When the ASVAB was introduced in January 1976, with the inflated scale, enlistment standards were inadvertently lowered. Many applicants previously unqualified for service suddenly passed the test. As recruiters discovered that fewer people were failing the enlistment tests

* Note that the increase in category V for Army applicants, from 9 percent in December 1975 to 16 percent in January 1976, is a ratio of 1 to 1.7. This increase was used in chapter 2 to help estimate the amount of test compromise in the Army calibration sample.

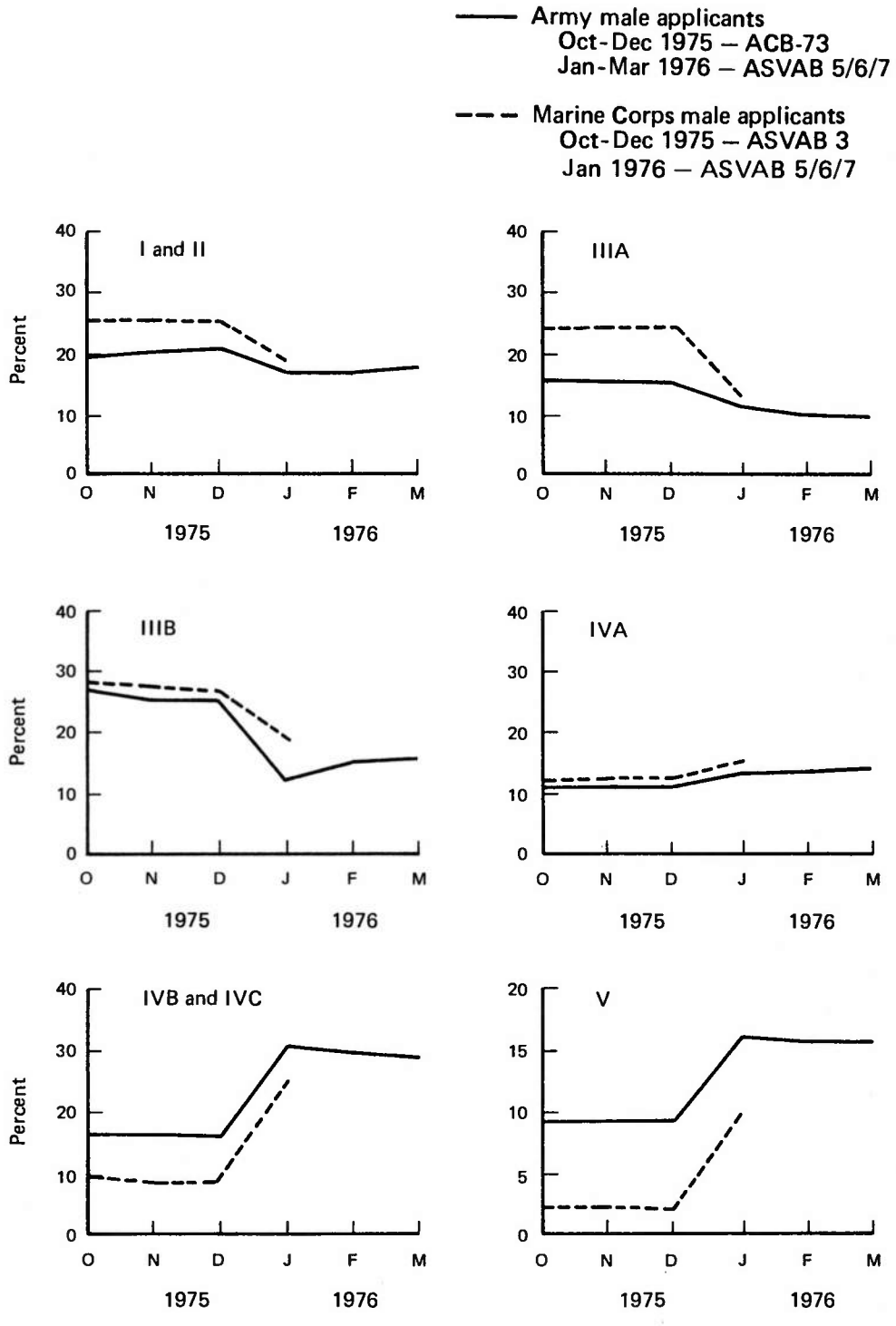

FIG. 9: SCORE DISTRIBUTIONS OF ARMY AND MARINE CORPS MALE APPLICANTS, CORRECT ASVAB 5/6/7 SCORE SCALE

they could have recruited applicants known to have low test scores from prior testing or those thought to have low scores based on other information, such as employment and school history. People with low employment prospects in the civilian economy are easier to recruit, and recruiters could more easily meet their goals by recruiting these types of people; typically, they score low on the ASVAB or they fail to qualify for enlistment. With the inflated scale, however, they qualified for enlistment. As some recruiters became aware of this fact, they may have taken the easy way and done their recruiting of people who are easy to recruit; that is, those marginally qualified. If recruiters responded in this way, then the percentage of marginally qualified persons in categories IIIB and IVA should increase during the month of January. The effects would likely be most pronounced for Army and Marine Corps applicants.

We computed the percentage of applicants in categories III, IV, and V, based on the original scale, for each processing day during December 1975 and January 1976. The percentages for Army applicants are shown in figure 10. The percentages for Marine Corps applicants are parallel to those for the Army.

Although the percentages in each category fluctuate from day to day, for each month they tend to be flat. The percentage in category V during January shows a slight downward trend, while that in categories IIIA and IIIB moves upward. The differences from the beginning to the end of the month, however, do not support the conjecture that the percentage of applicants with low ability increased during the month. The increase in category V was immediate and remained at about the same level during the month. Based on the lack of change in scores during the month of January 1976, we concluded that the ability of applicants did not change when the new test was introduced. We now examine the second assumption that scores on the tests used in 1975 are accurate.

ARE SCORES ON THE OLD TEST ACCURATE?

Concern about the accuracy of ACB-73 and ASVAB 3 scores has been a persistent issue throughout the analysis. In chapter 2 we focused on the accuracy of the score scales themselves or on the effects of coaching. In this chapter we consider a new scource of possible error--accuracy of scoring the ASVAB 3 and ACB-73 answer sheets. We then consider the combined effects of three possible sources of error--incorrect scoring procedures, test compromise, and inflated score scales--to help explain the large decrease in AFQT test scores if the correct scale had been used in January 1976. This discussion will also help explain why inflation of the original ASVAB 5/6/7 scale in the category III range was not apparent in the distributions of operational or observed test scores.

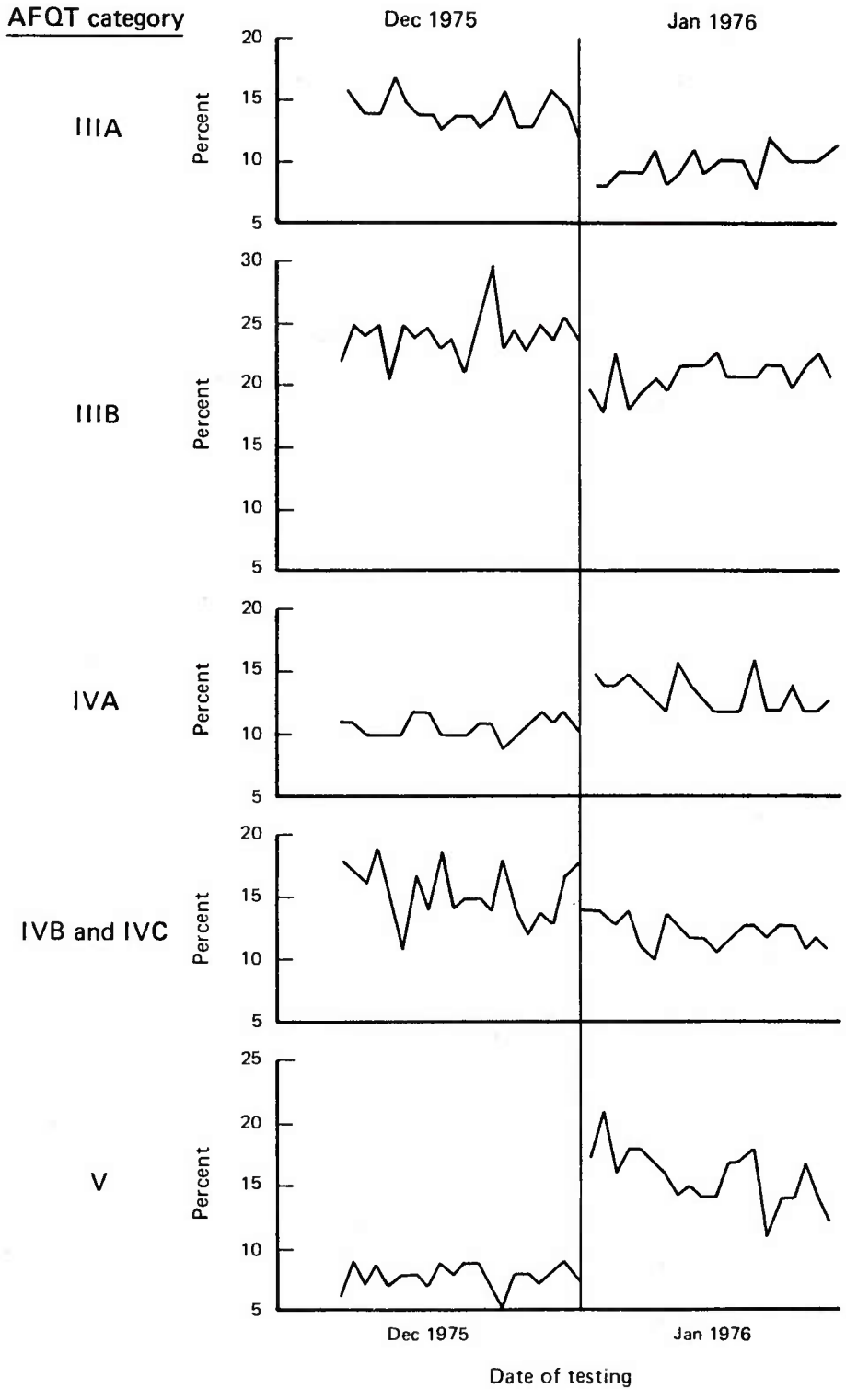

FIG. 10: PERCENT OF ARMY MALE APPLICANTS IN AFQT CATEGORIES III, IV, AND V, SHOWN FOR EACH DAY OF TESTING IN DEC 1975 AND JAN 1976

Were the Old Tests Scored Accurately?

To try to answer the question, we need to review the variety of test batteries and scoring formulas that were used in 1975. These include:

- ACB-73--scored rights only (number of items answered correctly)
- ASVAB 3--first subtest scored rights only and the remaining subtests scored rights minus one-third the number of wrongs (R - W/3) (items omitted were not scored as either right or wrong)
- Navy Basic Test Battery (NBTB)--scored rights only.

The ACB-73 was administered to all Army applicants; ASVAB 3, to all Marine Corps applicants and to Air Force applicants who had already passed rigorous screening at Air Force recruiting stations; and the Navy Basic Test Battery to Navy applicants who had passed rigorous screening at Navy recruiting stations.

At that time the AFEES used two scoring procedures. The smaller stations scored the answer sheets by hand. No systematic error is expected from hand scoring, although hand scoring frequently produces random errors. The larger stations used optical scanners to score the answer sheets. The scanners had a scoring formula switch that determined how wrong answers would be counted. If the scoring switch were not moved when scoring the different test batteries, then the scores would be in error. The most likely error is that ASVAB 3 would have been scored as rights only and the Marine Corps and Air Force examinees would be placed in higher AFQT categories than they should be. Because ACB-73 and the Navy Basic Test Battery are scored rights only, they probably were done correctly.

Incorrect scoring of ASVAB 3, to the extent it occurred, works in the same direction as test compromise and inflated score scales. The scores of some examinees would be unduly high, and the effect would be to help mask inflation of the ASVAB 5/6/7 scale.

Does Coaching Help Explain the Decrease in Test Scores?

In these paragraphs we discuss how coaching on the tests used in 1975 could explain the decrease in scores that would have occurred if the correct scale had been used. For coaching to explain the decrease in the upper half of the score range, we need to have a rationale for its extension. Also, we need to discuss why our earlier analysis of Army examinees in the 1975 calibration sample indicated that compromise was limited to category IIIB and below. As shown earlier in figure 5, the ASVAB 5/6/7 score scale computed for Army examinees tested in September through October 1975 was inflated above a raw score of 46,

which corresponds to a percentile score of 50 in the correct scale. This inflation suggests that the ACB-73 reference test scores were inflated throughout the score range by test compromise.

Coaching on ACB-73 and ASVAB 3 could have occurred in 1975 at the higher levels because enlistment guarantees were more attractive for individuals with AFQT scores of 50 and above. Recruits with above-average AFQT scores could qualify for more desirable skills, such as electronics technicians, and they were eligible to receive cash bonuses. Coaching could very well have moved a sizable percentage of recruits from categories IIIB and IVA into IIIA and above.

The PAFQT, especially for Army recruits, would not be an effective way to detect compromise at the higher levels. Because guarantees and bonuses were related to specific MOS groups, recruiters probably coached applicants on all relevant subtests. A likely technique is that booklets containing sample test items would be shown to applicants. The sample booklets could include all the items in the operational tests. With a little effort, an applicant of average ability could learn the answers to most of the questions, and qualify for a desired skill. Applicants at the low end of the scale would more likely be content with qualifying for enlistment, with the main emphasis on raising their AFQT scores. Thus, the PAFQT would be more sensitive in identifying suspected compromise in the lowest two categories whereas compromise on all subtests could occur in the upper categories.

The decreased percentage of applicants in the above-average range is plausible. Recruiters and applicants would have an interest in raising test scores above the minimum qualifying levels for bonuses and guarantees. When a new test is introduced, coaching would be reduced until the items become generally available among recruiters. Compromise at the low end, out of categories IV and V, would explain the sharp increase in these categories in January 1976 if the correct scale had been introduced. The shift in scores throughout the entire score range, therefore, is plausible.

INFLATED SCORES IN 1975 MASKED INFLATED SCORES IN 1976

The score distributions for January 1976 shown in figures 8 (the observed distributions based on the original scale) and figure 9 (the distribution if the correct scale been introduced) lead to different conclusions about the accuracy of the scores in 1975. A reasonable conclusion from examining figure 8 is that both the 1975 scores and the original score scale are accurate in categories V, IV, and IIIB. The ASVAB Working Group in 1976 reached essentially that conclusion. But from figure 9, the 1975 scores appear to be faulty. Our analysis supports the fact that the 1975 scores were inflated throughout the score range by coaching, particularly in category IVB and below. The amount of coaching we found in chapter 2, that the number of examinees in categories IVB, IVC, and V should be increased by 70 percent, accounts quite well for the small shift of observed scores in categories

IVB and IVC betweeen December 1975 and January 1976 (figure 8). At the upper end, the amount of coaching was smaller, and it did not mask the inflation of the original ASVAB 5/6/7 scale.

In 1975 and 1976 the amount of coaching was not known. At that time the PAFQT was not in use, and the ASVAB Working Group had no tool for estimating the effect of coaching on the AFQT score distributions. Only through the development of the PAFQT are we now able to compute how much coaching may be affecting the AFQT scores.

Our conclusion is that inflated scores in 1975 did mask the inflation of the 1976 scores. In 1975 the primary reason for inflated scores was coaching, although inflated scales for ACB-73 and ASVAB 3 may have had a small effect in the category IVA range. Incorrect scoring of ASVAB 3 answer sheets may also have contributed to inflation of the 1975 scores. An outcome of the masking effect was that the ASVAB Working Group judged that no correction to the ASVAB 5/6/7 score scale in the low range was required.

The assumption that scores on the old test are accurate is not tenable. Because of coaching, possible scoring errors, and some inaccuracy of the scales for ACB-73 and ASVAB 3, the 1975 scores were themselves inflated, and they helped mask inflation of the ASVAB 5/6/7 scale at the low end. We now turn to the question about the accuracy of the 1976 scores.

ARE SCORES FOR ASVAB 5/6/7 ACCURATE?

The accuracy of the ASVAB 5/6/7 scores may be affected by incorrect scoring of the AFQT in January 1976 and of course by the now familiar refrain that the correct scale may be too difficult. First, we examine the accuracy of scoring the ASVAB 5/6/7 answer sheets, and then we discuss what we have to believe about coaching to believe that the correct scale is accurate.

Was ASVAB 5/6/7 Scored Correctly?

When ASVAB 5/6/7 was introduced, AFEES personnel had almost no time to learn how to administer and score it. Scoring errors could have occurred, resulting in lowered AFQT scores. Available evidence indicates that the tests were scored correctly. The score distributions of examinees known to have taken ASVAB 6 or 7 in early 1976 agree closely with the distributions for examinees known to take other test forms during this period. Details of this analysis are in appendix I. Further supporting evidence can be inferred from figure 8. The score distributions for January, February, and March 1976 are similar, which suggests that the scoring procedures were similar throughout the period. Three months is more than enough time for AFEES personnel to become proficient in scoring the new test. The ASVAB 5/6/7 answer sheets appear to have been scored correctly.

We concluded that the ASVAB 5/6/7 scores in January 1976, based on the correct score scale, accurately measure the aptitudes of applicants.

Coaching and the Credibility of the Correct ASVAB 5/6/7 Scale

The role of coaching has been crucial throughout our analysis to establish the credibility of the correct ASVAB 5/6/7 scale. First, we used it to help reproduce the correct scale in the original ASVAB 5/6/7 calibration sample. We estimated that the effects of coaching contributed up to 6 percentile score points to the inflated scale. Second, we concluded that it was the primary reason the test scores in 1975 were inflated, which had the effect of masking inflation of the original ASVAB 5/6/7 scale. Third, we made an implicit assumption about coaching when we compared the AFQT distributions in September 1980 based on the correct scale with the October 1980 distributions for ASVAB 8/9/10.

We assumed that the amount of coaching in September 1980 was negligible. When we compared the September and October 1980 AFQT distributions, we noted they were similar and concluded that the correct scale is generally accurate. If, however, the ASVAB 5/6/7 scores in September 1980 were inflated by coaching, then they could not be compared directly to the ASVAB 8/9/10 scores in October. We believe our assumption is correct because the use of PAFQT in 1980 helped identify examinees who were coached on the AFQT. In addition, all services placed command emphasis on identifying and eliminating recruiter malpractice. If, however, coaching was extensive in September 1980, then the effects would be to mask any tendency for the correct scale to be too difficult. Our assumption about negligible compromise in 1980 appears plausible, and we see no reason to change our conclusion that the correct scale is generally accurate.

CHAPTER 5

DISCUSSION AND CONCLUSIONS

UNCERTAINTY ABOUT THE ASVAB SCORE SCALE

The attempt to unravel what went wrong with the original calibration of ASVAB 5/6/7 has led us down avenues we had not suspected we would travel when we began our analysis. We found three plausible reasons that explained the error: likely incorrect scoring of the reference test for the Navy and Air Force recruit samples, coaching on the reference test used for Army examinees, and selection of Army examinees based on their operational reference test scores.

Our analysis of coaching in the Army sample has an objective basis, but there is no way to verify that the amount is what we estimated it to be. An element of uncertainty remains about the amount of coaching. In addition, we could not exactly reproduce the correct scale in the original calibration sample, and we are left with more uncertainty about the accuracy of the correct scale.

When we attempted to verify the accuracy of the ACB-73 and ASVAB 2/3 scales, we encountered more uncertainty. We found, as discussed in appendix D, that the ACB-73 scale may be inflated in the category IV range; and, in addition, we noted the score scales for ACB-61, widely used by the Army from 1959 until 1973 and by the Marine Corps from about the same time until 1976, were themselves uncertain.

The score distributions for AFQT 7/8 also are filled with uncertainty. Although the scale for AFQT 7/8 is accurate, the scores during the AVF era appear to be inflated by test compromise.

Wherever we turned for a stable anchor point for evaluating the accuracy of the correct ASVAB 5/6/7 scale we encountered uncertainty. One conclusion from all this uncertainty is that the score distributions during the AVF era cannot be taken at face value. We cannot assert with confidence how many accessions were in category IV or in any other category. Because of coaching and different score scales, we also doubt whether changes in score distribution across time can be interpreted as reflecting changes in aptitudes of the examinees; other variables, such as coaching, could be the reason. A new ASVAB score scale based on a new reference population would remove uncertainty about the scale. A new reference population for ASVAB 8/9/10 is available from the 1980 youth population. ASVAB 8 was administered to a nationally representative sample from the current youth population, and a new scale can be computed from the data. At the time of writing plans are underway to develop a new ASVAB scale.

Effects on DoD Testing Program

One significant outcome of the miscalibration is that Congress directed that a DoD Testing Advisory Committee be established to provide guidance on the military selection and classification testing program. The Committee's initial efforts were to review both the calibration of ASVAB 8/9/10, which the members agreed was done correctly, and the research evidence on the predictive validity of the ASVAB.

A second significant outcome is that each service has been required to validate its enlistment standards against job performance. The inflated ASVAB 5/6/7 scale in effect lowered enlistment standards below their nominal level, resulting in an influx of low-scoring recruits. A natural question is how job performance was affected by these lowered standards. During 1980, each service initiated research efforts to determine the feasibility of validating enlistment standards against job performance.

The error in the ASVAB score scale also has had a dramatic effect on the procedures employed by DoD for calibrating new tests. In 1975, shortcuts were used to calibrate ASVAB 5/6/7 because of constraints on extra testing at AFEES. When the magnitude of the error in the scale became apparent, personnel management recognized the importance of following established procedures for calibrating new tests. When ASVAB 8 was calibrated in 1980, the required resources were made available to conduct the study properly. During its first year of use, there were no complaints that the scale appeared to be in error. In fact, a follow-on evaluation demonstrated conclusively that the calibration had been done correctly. Through this study to calibrate ASVAB 8, DoD learned that AFEES can handle the extra burden and that applicants are not turned away. Subsequent calibration efforts have been conducted at AFEES with little trouble.

SUMMARY

Our objectives were to find what went wrong in the original calibration of ASVAB 5/6/7, to reproduce the correct scale in the original calibration sample, and to discuss problems with interpreting distributions of operational scores in the AVF era.

Through reanalysis of the original data collected in 1975 and of related data, we have reproduced the correct ASVAB score scale within 2 to 3 raw score points. The incorrect scoring of the ASVAB 2 reference tests administered to Navy and Air Force recruits produced an inflated scale in categories I, II, and IIIA. By adjusting the ACB-73 scores for test compromise and the scale for prior selection of examinees in the Army sample, we reproduced the correct scale in categories IVB, IVC, and V. In categories IIIB and IVA the inflation was reduced but persisted in all samples--adjusted Army sample, and Navy and Air Force recruits with the ASVAB 2 reference test scored correctly. The correct scale is accurate in this range, however, as evidenced by the multiple studies

that are in essential agreement. Our first two objectives were attained reasonably well.

There are serious problems with interpreting operational scores in the AVF era. Our analysis of coaching also explains why the inflated scale in categories IVB and IVC was not apparent when comparing the December 1975 and January 1976 distributions. Coaching, coupled with some inflation of ACB-73 and ASVAB 3 scales and perhaps some incorrect scoring of ASVAB 3 answer sheets, raises serious doubts about the meaning of score distributions throughout the AVF era. The adoption of a new reference population, with a new score scale, should remove the uncertainty that has surrounded the ASVAB.

REFERENCES

[1] CNA, Study 1152,"A Reexamination of the Normalization of the Armed Services Vocational Aptitude Battery (ASVAB) Forms 6, 7, 6E, and 7E," by William H. Sims and Ann R. Truss, Unclassified, Apr 1980

[2] Office of the Secretary of Defense (Directorate for Accession Policy), Technical Memorandum 80-1, "Renorming ASVAB 6/7 at Armed Forces Examining and Entrance Stations," M. H. Maier and F. C. Grafton, Unclassified, Aug 1980

[3] Office of the Secretary of Defense (Directorate for Accession Policy), Technical Memorandum 80-2 "Scaling of the Armed Services Vocational Aptitude Battery, Form 7, and the General Classification Test to the Armed Forces Qualification Test Scale," R. F. Boldt, Unclassified, Aug 1980

[4] Office of the Secretary of Defense (Manpower, Reserve Affairs, and Logistics), "A Review and Analysis of Score Calibration for the Armed Services Vocational Aptitude Battery, Appendix B of "Aptitude Testing of Recruits, A Report to the House Committee on Armed Services," by Richard M. Jaeger, Robert L. Linn, and Melvin R. Novick, Unclassified, Jul 1980

[5] Office of the Secretary of Defense (Directorate for Accession Policy), "History of the Armed Services Vocational Aptitude Battery (ASVAB) 1974-1980," ASVAB Working Group, Unclassified, Mar 1980

[6] Air Forces Human Resources Laboratory, AFHRL-TR-76-87, "Development of the Armed Services Vocational Aptitude Battery, Forms 5, 6, and 7," by H.E. Jensen, I.H. Massey, and L.D. Valentine, Jr., Unclassified, Dec 1976

[7] Department of Defense, DoD 1304.12W, Conversion Tables for Armed Services Vocational Aptitude Battery (ASVAB) Forms 6-7, Unclassified, Jan 1976

APPENDIX A

CALIBRATION OF ASVAB 5/6/7 ON SAMPLES OF SERVICE RECRUITS IN SPRING 1976

APPENDIX A

CALIBRATION OF ASVAB 5/6/7 ON SAMPLES OF SERVICE RECRUITS IN SPRING 1976

The Navy, Air Force, and Marine Corps collected ASVAB 5/6/7 and reference test scores on samples of recruits in spring 1976. The Navy tested two samples of recruits. One sample of 240 recruits was tested with ASVAB 6 or 7 at the time of enlistment or on the fifth day of processing at the San Diego Recruit Training Center (RTC). They were tested with AFQT 7 on the sixth day of processing at the RTC. A second sample of 293 Navy recruits was also tested with ASVAB 6 or 7 at the time of enlistment or on the fifth day of processing at the San Diego RTC. However, the reference test, AFQT 7A, was not administered until the sixth week of boot camp. The reference test for both Navy samples was administered in April and May 1976. The Air Force sample of 352 recruits was administered ASVAB 6 and AFQT 7A on 5 May 1976.

The sample of Marine Corps recruits is more difficult to describe. Results made available to the ASVAB Working Group in June 1976 were based on a sample of 930 Marine recruits. This sample was tested with both ASVAB 6 or 7 and the Army Classification Battery (ACB-61), which at that time was routinely administered at Marine reception centers to all recruits. The sample of 930 recruits may be part of a larger sample tested from December 1975 until March 1976 with both ACB-61 and ASVAB 5/6/7. During this period about 8,000 recruits were tested. The Marine Corps sample used three subtests (Word Knowledge, Arithmetic Reasoning, and Spatial Perception) from ACB-61 as the reference test.

The equipercentile equating technique was used to calibrate ASVAB 5/6/7 to AFQT 7 for the two Navy samples and the Air Force sample. The results are shown in table A-1. Also shown in table A-1 are the calibrations based on the stratified sample of 930 Marine recruits, using ACB-61 as the reference test, and the original and correct ASVAB 5/6/7 score scales. Because the Navy and Air Force sample had few cases with low test scores, the score scale could not be computed reliably below a percentile score of 20. Above this point, the scoring on the Navy and Air Force recruits generally agrees with what in 1980 proved to be the correct score scale. The agreement with the correct score scale is much closer than with the original 1975 scale.

TABLE A-1

CALIBRATION OF ASVAB 5/6/7
FOR SERVICE RECRUITS IN SPRING 1976

Raw score	Percentile score						Raw score	Percentile score					
	Navy[a]	Navy[b]	AF[c]	MC[d]	Original	Correct		Navy	Navy	AF	MC	Original	Correct
0-12					1		47	53	50	49	63	69	53
13					2		48	55	52	50	67	71	56
14					2		49	57	54	51	69	74	58
15				1	2		50	60	57	53	70	75	60
16					3	1	51	62	60	57	73	77	62
17					3	2	52	63	63	59	74	79	65
18				3	3	3	53	65	65	62	76	80	67
19				3	4	4	54	68	66	65	78	82	70
20				4	6	5	55	72	69	68	80	84	72
21				5	7	6	56	75	71	71	83	85	75
22				6	8	7	57	78	75	73	84	87	77
23				7	9	9	58	80	78	75	86	88	80
24				9	11	10	59	82	81	77	87	89	82
25				11	13	11	60	84	84	80	88	92	84
26	11			14	17	12	61	85	87	83	90	93	86
27				16	18	13	62	86	89	85	91	94	87
28	13	13		16	21	14	63	88	90	87	92	95	89
29				21	21	15	64	89		88	95	96	91
30	15	14	15	23	25	16	65	91		90	96	97	93
31			16	25	30	17	66	94		92	98	98	95
32	17	18	18	28	33	18	67	97		94		98	97
33	18	21	19	31	36	19	68			96	99	99	98

TABLE A-1 (Cont'd)

Raw score	Percentile score						Raw score	Percentile score					
	Navy[a]	Navy[b]	AF[c]	MC[d]	Original	Correct		Navy	Navy	AF	MC	Original	Correct
34	21	22	20	33	38	21	69			98		99	99
35	24	25	22	35	42	23	70			99		99	99
36	27	25	23	38	43	25							
37	30	26	25	40	45	27							
38	32	28	27	42	48	29							
39	35	30	29	43	49	31							
40	38	33	31	45	54	33							
41	42	36	33	48	58	35							
42	44	39	34	51	60	38							
43	47	41	36	53	62	41							
44	49	42	38	56	64	44							
45	50	45	40	58	65	47							
46	51	48	46	61	67	50							

	Sample	Reference test
[a]293	Navy recruits.	AFQT 7A
[b]240	Navy recruits.	AFQT 7A
[c]352	Air Force recruits.	AFQT 7A
[d]930	Marine Corps recruits.	ACB-61

The results for the Marine Corps sample deviate from all other calibrations. These results indicate that the original score scale is inflated throughout the range, but not as much as indicated by the Navy and Air Force samples or as shown by the correct score scale. Essentially, these same results appear in [A-1].*

* The operational score scale adopted by the ASVAB Working Group in July 1976 and implemented September 1976 does not agree with any of these scales. The Working Group appears to have adopted the scale based on Air Force and Navy recruits in the top half of the range, where the results are most reliable, but not in the bottom half, where they may be less reliable. The discrepant results for the Marine Corps recruits, together with the original score scale, may have influenced the Working Group to leave the bottom half of the score virtually intact.

REFERENCES

[A-1] Office of the Secretary of Defense (Directorate of Accession Policy), "History of Armed Services Vocational Aptitude Battery (ASVAB) 1974-1980," ASVAB Working Group, Unclassified, Mar 1980

APPENDIX B

DEVELOPING AND INTERPRETING THE ASVAB SCORE SCALE

APPENDIX B

DEVELOPING AND INTERPRETING THE ASVAB SCORE SCALE

DEVELOPING THE ASVAB SCORE SCALE

Scores on the ASVAB and AFQT are intended to provide historical continuity in the measurement of ability of the enlisted force. The AFQT score categories were developed on all officers and enlisted men serving under arms during World War II. Scores on new forms of the AFQT and on ASVAB have been related to the original population using statistical equating procedures.

Equating Tests to a Reference Population

Tests that have the same score scale show the same level of ability or aptitude relative to a reference population. The technique commonly used to put tests on the same scale is called "equipercentile equating."

For simplicity let us start with one test for which we want to build a score scale. For all ability or aptitude tests, such as the ASVAB, the number of items correct, called the raw score, does not show directly how able an examinee is. The raw score must be referenced to a standard, and the standard used for aptitude tests is relative standing in a meaningful population of persons. The population must have some interest to the test user. For example, a college admissions officer or school counselor may be interested in how an examinee compares to all college applicants, or all high school seniors, or applicants for a particular college. Personnel managers in the services may be interested in how applicants for enlistment compare to the current population, to all other applicants in a given time period, or historically over the years. (More will be said on these possible reference groups later on.) Relative standing in a population is usually expressed as a percentile score that ranges from 1 (low) through 100 (high), with 50 as average. Each percentile score is equal to 1 percent of the population. Thus, a percentile score of 30 means that 30 percent of the population score at or below that point, and 70 percent score above. Raw scores on a test are converted to percentile scores on the basis of how many examinees in a sample representative of the population score above or below each raw score. The resulting conversion from raw score to percentile score is called the "test norms," and the process of determining how to convert raw scores is called "normalization," "norming," or sometimes "standardization."

Norming a single test is a simple straightforward procedure: administer the test to a representative sample from that population; count how many examinees score at or below each raw score; and then convert the raw scores to percentile scores, where each percentile score

is 1 percent of the population, and the scores range from 1 through 100. The sample must be carefully drawn to make sure it is representative of the population, and the test administration must be done carefully to ensure the scores are accurate, but the computations to norm the test are basically simple counting.

Now, let us complicate matters a bit. Suppose we want to develop a new form of the test, but still keep the same norms, or score scale, we had for the old test. When new forms are introduced every few years, or even more frequently, as with college admissions tests, the scale should not fluctuate with each new form. If the scales do fluctuate, then personnel managers do not know what a percentile score means in terms of expected performance. The process of relating the scores on a new test to scores on an old test is called "scaling" or "calibration."

Conceptually, scaling a new test to an old one is also straightforward. We start with the fact that we have norms for the old test. One common procedure is to administer both the old and new test to a sample from the reference population, which for the ASVAB is the World War II mobilization population. Raw scores on the new test are set equal to the percentile scores on the old test by using the equipercentile equating technique. The equipercentile equating technique consists of the following basic steps:

- Compute the cumulative frequency distribution of both the old or reference test and the new test that is to be calibrated to the old test.

- Plot the cumulative frequencies for both the reference and new tests on the same graph. (Note that the horizontal axis represents the scores for both the reference and new tests; the vertical axis represents the cumulative percentage of the sample for both tests.)

- Determine the scores on both the new and old test that correspond to the same cumulative percentage. Each score on the new test is converted to the percentile score on the reference test that has the same cumulative frequency. The equipercentile equating technique is used in appendix C.

INTERPRETING THE ASVAB SCORE SCALE

Until the problem with ASVAB 5/6/7 score scale surfaced, procedures for developing ASVAB score scales were of interest only to the small band of testing psychologists employed by the Department of Defense. A common assumption of test users was that the percentile scores referred to the current population, either of mobilization age or of applicants for enlistment. Thus, a percentile score of 30 on AFQT was frequently thought to refer to relative standing in some unknown but current group,

and certainly not to some population tested more than 35 years ago. As the facts about the reference population became more widely known, the criticism also grew. The criticism is justified from one point of view, but not from another, and a few statements on the merits of different types of scale norms may be illuminating for some test users.

Enlistment standards, as for most employment and admissions standards, are based on some combination of performance requirements (quality) and personnel supply (quantity). An institution, whether military service, college, or industrial concern, places certain demands on its members. Tests are used to help determine which individuals are likely to satisfy the institutional requirements and which individuals are likely to fail. If the potential personnel supply were practically inexhaustible, then the institution could raise entrance standards to the point where almost everyone accepted would be satisfactory. Because very few institutions can be so highly selective, especially considering a social equity point of view, as during periods of induction, the setting of standards usually also involves considerations of personnel supply. Institutional requirements in the military are not absolutely fixed; instead they can be varied depending on the quality of accessions. Thus, both quality requirements and availability of potential accessions influence standards.

The two considerations in setting standards, quality and quantity, each requires a different kind of score scale. Questions about quality are best addressed by a score scale that has stability across different tests and different periods of time; hence, the decision to keep the AFQT and ASVAB scales referenced to the same population dating to World War II. The particular reference population is largely irrelevant. The important consideration is that over the years, test users gain a stable, accurate expectation about the performance level of individuals at the various score levels. Persons with lower scores can be expected to be reasonably satisfactory in some types of training programs and jobs, but not in others. As long as the expected performance of persons with the same level of scores remains relatively constant, then personnel managers can maximize the utility of their decisions. But if the meaning of the scores changed whenever a new test form is introduced, then users would be in a quandary about the level of performance to expect for each new group of accessions. A stable score scale is especially useful to trainers and others concerned about quality.

Score scales that refer to a population of yesteryear have relatively little meaning, however, for personnel managers with primary responsibility for the quantity or supply of accessions. A reasonable concern about quantity is how many and what sort of people are entering relative to the potential supply. This kind of concern can best be addressed by a scale that refers to a current population. The same scale cannot address both the quality and quantity for any extended length of time; one or the other interpretation becomes out of date.

Inflation of the ASVAB score scale refers only to the question of how scores on a particular form are calibrated to the traditional reference population. The scale does not show how the current youth population would score on the tests. The decision until now has been to keep the ASVAB referenced only to the stable score scale, and not to reference it to the current population. In the future we expect to be able to reference ASVAB scores to both the traditional population (to address concerns about quality) and the current youth population (to address concerns about quantity).

APPENDIX C

CUMULATIVE FREQUENCY DISTRIBUTIONS OF REFERENCE TEST AND ASVAB 5/6/7 SCORES

APPENDIX C

CUMULATIVE FREQUENCY DISTRIBUTIONS OF REFERENCE TEST AND ASVAB 5/6/7 SCORES

The cumulative frequencies of the reference test percentile scores and ASVAB 5/6/7 raw scores are presented in this appendix. Table C-1 presents the cumulative frequencies for the full sample of males (N = 4,588) and table C-2 for the females (N = 558). The ASVAB 2 reference test scores for Navy and Air Force recruits are the original ones from the data tape provided by the Air Force Human Resources Laboratory (AFHRL).

The equipercentile equating technique was used to calibrate raw scores on ASVAB 5/6/7 to the reference tests. The technique equates scores attained by the same cumulative percent of the sample. For example, in table C-1, 29 percent of the sample scored 36 or below on the reference test, also, 29 percent of the sample scored 32 or below on ASVAB 5/6/7. The ASVAB 5/6/7 raw score of 32 therefore would be set equal to a reference test score of 36. The reference test in table C-1 is expressed as percentile scores. The ASVAB 5/6/7 in table C-1 is expressed as raw scores. In the Department of Defense, percentile scores are defined as the percentage of the reference population that scores at or below that point. ASVAB 5/6/7 raw scores are the number of items correct. Unless specified otherwise, we used the equipercentile equating technique throughout this report for constructing score scales.

The cumulative frequencies of the reference test scores for Navy and Air Force recruits (ASVAB 2) and ASVAB 5/6/7 raw scores are shown in table C-3 and C-4, respectively.

The reference test scores for the Navy and Air Force recruits are shown both as originally computed by the AFHRL and as rescored by us using the scoring rule: number of right answers minus one-third the number of wrong answers, counting all omitted items as wrong. The rescored reference test distributions were used to compute new score scales for ASVAB 5/6/7 using the equipercentile equating technique.

The cumulative distributions for the Army sample, limited to examinees tested in September and October 1985, are shown in table C-5. The ACB-73 distribution is based on the scores of record obtained from the Defense Manpower Data Center (DMDC) rather than the reference test scores coded on the ASVAB 5/6/7 answer sheets. The conversion from ASVAB 5/6/7 raw score to percentile score for this Army sample is also shown in table C-5.

Table C-6 shows cumulative distributions for the combined sample of Army examinees tested in September and October 1975 and Navy and Air Force recruits. The reference test scores for the Navy and Air Force

recruits were rescored to adjust for the estimated number of wrong answers, and the reference test scores for the Army examinees were adjusted for test compromise. The ASVAB 5/6/7 raw scores were not adjusted.

TABLE C-1

CUMULATIVE FREQUENCIES FOR ORIGINAL ASVAB 5/6/7
CALIBRATION SAMPLE, MALES[a]

Reference test[b]		ASVAB 5/6/7	
Percentile score	Percent[c]	Raw score	Percent[c]
1	0	0-12	1
2	0	13	1
3	1	14	1
4	1	15	1
5	1	16	2
6	2	17	2
7	2	18	3
8	3	19	4
9	4	20	5
10	4	21	7
11	4	22	8
12	6	23	9
13	6	24	11
14	7	25	13
15	7	26	15
16	8	27	17
17	8	28	19
18	10	29	21
19	10	30	23
20	11	31	26
21	13	32	29
22	13	33	31
23	13	34	34
24	13	35	36
25	15	36	39
26	15	37	42
27	16	38	44
28	18	39	46
29	18	40	49
30	18	41	52
31	22	42	54
32	22	43	57
33	24	44	59
34	25	45	62
35	26	46	64
36	29	47	66
37	30	48	69
38	32	49	71
39	34	50	73
40	34	51	75

TABLE C-1 (Cont'd)

Reference test[b]		ASVAB 5/6/7	
Percentile score	Percent[c]	Raw score	Percent[c]
41	36	52	77
42	37	53	79
43	38	54	81
44	41	55	83
45	41	56	85
46	42	57	87
47	45	58	89
48	47	59	90
49	47	60	92
50	50	61	93
51	50	62	84
52	51	63	95
53	52	64	96
54	52	65	98
55	53	66	98
56	53	67	99
57	54	68	100
58	54	69	100
59	56	70	100
60	56		
61	58		
62	59		
63	61		
64	61		
65	63		
66	63		
67	63		
68	66		
69	66		
70	69		
71	69		
72	72		
73	72		
74	75		
75	75		
76	75		
77	75		
78	77		
79	77		
80	80		
81	80		
82	82		
83	82		

TABLE C-1 (Cont'd)

Reference test[b]		ASVAB 5/6/7	
Percentile score	Percent[c]	Raw score	Percent[c]
84	85		
85	85		
86	87		
87	87		
88	88		
89	88		
90	90		
91	92		
92	94		
93	95		
94	96		
95	97		
96	98		
97-99	100		

[a]N = 4,588.
[b]Reference test is ASVAB 2 for Navy and Air Force recruits, original scores, and ACB-73 for Army sample.
[c]Cumulative percent.

TABLE C-2

CUMULATIVE FREQUENCIES FOR ORIGINAL ASVAB 5/6/7
CALIBRATION SAMPLE, FEMALES[a]

Reference test[b]		ASVAB 5/6/7	
Percentile score	Percent[c]	Raw score	Percent[c]
1-10	1	0-18	1
11	2	19	1
12	2	20	2
13	3	21	3
14	3	22	3
15	3	23	4
16	4	24	5
17	4	25	6
18	4	26	7
19	4	27	9
20	4	28	10
21	4	29	11
22	4	30	12
23	5	31	12
24	5	32	13
25	7	33	15
26	7	34	16
27	7	35	18
28	8	36	22
29	8	37	23
30	8	38	25
31	10	39	28
32	10	40	31
33	11	41	33
34	12	42	34
35	13	43	37
36	13	44	40
37	14	45	43
38	15	46	45
39	15	47	49
40	15	48	51
41	15	49	55
42	17	50	58
43	18	51	62
44	19	52	64
45	19	53	67
46	22	54	71
47	22	55	73
48	24	56	75
49	24	57	78

TABLE C-2 (Cont'd)

Reference test[b]		ASVAB 5/6/7	
Percentile score	Percent[c]	Raw score	Percent[c]
50	27	58	81
51	27	59	84
52	29	60	86
53	29	61	88
54	29	62	90
55	32	63	93
56	32	64	94
57	34	65	96
58	34	66	96
59	37	67	97
60	37	68	99
61	38	69	99
62	39	70	100
63	42		
64	42		
65	47		
66	47		
67	47		
68	50		
69	50		
70	52		
71	52		
72	52		
73	55		
74	55		
75	60		
76	60		
77	60		
78	63		
79	63		
80	66		
81	66		
82	72		
83	72		
84	75		
85	75		
86	77		
87	78		
88	81		
89	81		
90	84		
91	87		
92	89		

TABLE C-2 (Cont'd)

Reference test[b]		ASVAB 5/6/7	
Percentile score	Percent[c]	Raw score	Percent[c]
93	92		
94	94		
95	95		
96	97		
97	98		
98	99		
99	100		

[a]N = 558.
[b]Reference test is ASVAB 2 for Navy and Air Force recruits, and original scores and ACB-73 for Army sample.
[c]Cumulative percent.

TABLE C-3

CUMULATIVE FREQUENCIES OF REFERENCE TEST, ASVAB 2, AND ASVAB 5/6/7 SCORES, NAVY RECRUITS[a]

Reference test			ASVAB 5/6/7	
	Cumulative percent			
Percentile score	Original[b]	Rescored[c]	Raw score	Cumulative percent
1-6	1	1	0-22	1
7	1	2	23	2
8	1	2	24	2
9	1	2	25	3
10	1	2	26	3
11	1	2	27	3
12	1	2	28	4
13	1	3	29	5
14	1	3	30	6
15	1	3	31	7
16	1	3	32	8
17	1	3	33	9
18	1	4	34	11
19	1	5	35	14
20	1	5	36	16
21	1	5	37	18
22	1	5	38	19
23	2	6	39	22
24	2	6	40	25
25	2	8	41	28
26	2	9	42	31
27	2	9	43	34
28	2	9	44	38
29	2	9	45	41
30	2	11	46	44
31	3	14	47	48
32	3	14	48	51
33	3	16	49	55
34	3	16	50	58
35	3	16	51	61
36	3	16	52	65
37	4	19	53	69
38	4	19	54	72
39	5	23	55	74
40	5	23	56	77
41	5	23	57	79
42	5	26	58	82

TABLE C-3 (Cont'd)

Reference test			ASVAB 5/6/7	
	Cumulative percent			
Percentile score	Original[b]	Rescored[c]	Raw score	Cumulative percent
43	5	26	59	84
44	6	26	60	87
45	6	26	61	89
46	8	30	62	91
47	8	30	63	92
48	9	36	64	94
49	9	36	65	96
50	11	42	66	97
51	11	42	67	98
52	14	42	68	99
53	14	42	69	100
54	14	42	70	100
55	16	48		
56	16	48		
57	19	52		
58	19	52		
59	23	56		
60	23	56		
61	26	56		
62	26	56		
63	30	61		
64	30	61		
65	36	66		
66	36	66		
67	36	67		
68	42	70		
69	42	70		
70	48	70		
71	48	70		
72	48	70		
73	52	74		
74	52	74		
75	56	77		
76	56	77		
77	56	77		
78	61	80		
79	61	80		
80	66	80		
81	66	80		
82	70	84		
83	70	84		

TABLE C-3 (Cont'd)

Reference test		
	Cumulative percent	
Percentile score	Original[b]	Rescored[c]
84	74	87
85	74	87
86	77	89
87	77	89
88	80	89
89	80	89
90	84	91
91	87	93
92	89	95
93	91	93
94	93	97
95	95	97
96	97	99
97	99	99
98	100	100
99	100	100

[a]N = 1,086 males.
[b]ASVAB 2 scores as used in computing original ASVAB 5/6/7 score scale.
[c]ASVAB 2 rescored by subtracting one-third wrong answers.

TABLE C-4

CUMULATIVE FREQUENCIES OF REFERENCE TEST,
ASVAB 2 AND ASVAB 5/6/7 SCORES, AIR FORCE RECRUITS[a]

Reference test			ASVAB 5/6/7	
	Cumulative percent			
Percentile score	Original	Rescored	Raw score	Cumulative Percent
1-19	1	1	0-29	1
20	1	2	30	2
21	1	2	31	2
22	1	2	32	3
23	1	2	33	4
24	1	2	34	5
25	1	2	35	6
26	1	3	36	7
27	1	3	37	9
28	1	3	38	10
29	1	3	39	11
30	1	4	40	13
31	1	5	41	16
32	1	5	42	18
33	1	7	43	21
34	1	7	44	23
35	1	7	45	27
36	1	7	46	29
37	1	8	47	32
38	1	8	48	36
39	1	12	49	39
40	1	12	50	44
41	1	12	51	48
42	2	14	52	51
43	2	14	53	55
44	2	14	54	59
45	2	14	55	63
46	2	18	56	67
47	2	18	57	72
48	3	21	58	74
49	3	21	59	78
50	4	26	60	81

TABLE C-4 (Cont'd)

Reference test			ASVAB 5/6/7	
	Cumulative percent			
Percentile score	Original[b]	Rescored[c]	Raw score	Cumulative percent
51	4	26	61	84
52	5	26	62	88
53	5	26	63	90
54	5	26	64	92
55	7	32	65	95
56	7	32	66	97
57	8	36	67	98
58	8	36	68	99
59	12	42	69	100
60	12	42	70	100
61	14	42		
62	14	42		
63	18	46		
64	18	46		
65	21	51		
66	21	51		
67	21	51		
68	26	57		
69	26	57		
70	32	57		
71	32	57		
72	32	57		
73	36	62		
74	36	62		
75	42	67		
76	42	67		
77	42	67		
78	46	72		
79	46	72		
80	51	72		
81	51	72		
82	57	76		
83	57	76		
84	62	80		
85	62	80		
86	67	84		

TABLE C-4 (Cont'd)

Percentile score	Reference test: Cumulative percent — Original[b]	Reference test: Cumulative percent — Rescored[c]
87	67	84
88	72	84
89	72	84
90	76	88
91	80	91
92	84	94
93	88	94
94	91	96
95	94	96
96	96	99
97	99	99
98	100	100
99	100	100

[a]N = 990 males.
[b]ASVAB 2 scores as used in computing original ASVAB 5/6/7 scale.
[c]ASVAB 2 rescored by subtracting one-third wrong answers.

TABLE C-5

CUMULATIVE FREQUENCIES IN ARMY SAMPLE[a]

ACB-73 score of record[b]		ASVAB 5/6/7		
Percentile score	Cumulative percent	Raw score	Cumulative percent	Percentile score[c]
2		0-11		
3		12		
4		13	1	5
5	1	14	2	6
6	3	15	2	6
7	4	16	3	6
8	5	17	4	7
9	7	18	5	8
10	9	19	6	9
12	12	20	9	10
14	15	21	11	11
16	18	22	13	13
18	20	23	15	14
21	25	24	19	17
25	30	25	22	19
28	34	26	26	22
31	41	27	31	26
33	47	28	34	28
35	48	29	38	29
36	54	30	42	31
37	55	31	46	33
38	60	32	50	34
39	62	33	54	36
41	65	34	57	37
42	66	35	60	38
44	71	36	63	40
46	72	37	66	41
48	76	38	68	43
50	79	39	70	44
53	81	40	73	46
56	83	41	76	47
59	85	42	77	48
61	86	43	80	51
63	88	44	81	53
65	89	45	82	55
68	90	46	84	58
70	91	47	86	60
73	93	48	87	62

TABLE C-5 (Cont'd)

ACB-73 score of record[b]		ASVAB 5/6/7		
Percentile score	Cumulative percent	Raw score	Cumulative percent	Percentile score[c]
75	94	49	89	64
78	95	50	89	65
80	96	51	90	68
82	97	52	91	70
84	97	53	92	72
86	98	54	92	73
88	98	55	93	74
90	99	56	94	75
92-91	100	57	95	78
		58	96	80
		59	96	81
		60	97	82
		61	98	
		62-70	100	

[a]N is 1,151 cases tested with ACB-73 from September through October 1975.
[b]ACB-73 scores of record used as reference variable.
[c]ASVAB 5/6/7 percentile score computed on Army sample.

TABLE C-6

CUMULATIVE FREQUENCIES OF ADJUSTED REFERENCE TEST AND ASVAB 5/6/7 IN COMBINED CALIBRATION SAMPLE

Reference test				
	Cumulative percent		ASVAB 5/6/7	
Percentile score	A[a]	B[b]	Raw score	Cumulative percent
1-5	0	0	0-11	0
6	1	1	12	0
7	1	2	13	0
8	2	3	14	1
9	3	4	15	1
10	3	5	16	1
12	5	8	17	1
14	6	9	18	2
16	7	11	19	3
18	8	13	20	4
20	9	15	21	4
21	11	17	22	5
23	11	17	23	6
25	14	19	24	8
26	15	20	25	9
28	16	22	26	10
30	17	23	27	12
31	21	25	28	14
33	24	28	29	16
36	27	30	30	17
37	28	31	31	19
38	30	32	32	21
39	33	35	33	24
41	34	36	34	26
42	37	38	35	28
44	38	39	36	30
46	41	42	37	32
47	42	43	38	34
48	46	46	39	36
50	50	51	40	38
53	51	51	41	41
55	54	55	42	44
56	55	55	43	47
57	58	58	44	49
59	62	62	45	52
61	63	63	46	54
63	66	66	47	57
65	70	70	48	59
68	73	73	49	62

TABLE C-6 (Cont'd)

Percentile score	Cumulative percent		ASVAB 5/6/7	
	Aa	Bb	Raw score	Cumulative percent
71	74	74	50	65
73	77	77	51	68
75	80	80	52	70
78	82	82	53	73
80	83	83	54	75
82	86	86	55	78
84	88	88	56	80
86	90	90	57	83
88	91	91	58	85
90	93	93	59	88
91	94	94	60	89
92	96	96	61	91
94	97	97	62	92
96	99	99	63	94
97	99	99	64	95
98	100	100	65	97
99	100	100	66	98
			67	99
			68	99
			69	100
			70	100

[a]ASVAB 2 rescored for Navy and Air Force recruits, N = 2,076; ACB-73 scores of record.
[b]ACB-73 frequencies adjusted for test compromise.
[c]ASVAB 5/6/7 raw scores were not adjusted.

APPENDIX D

CALIBRATION OF ACB-73 AND ASVAB 2/3

APPENDIX D

CALIBRATION OF ACB-73 AND ASVAB 2/3

SUMMARY

The Army Classification Battery, form 1973 (ACB-73) and form 2 of the Armed Services Vocational Aptitude Battery (ASVAB 2) were used as the reference tests for calibrating forms 5, 6, and 7 of the ASVAB (ASVAB 5/6/7). In this appendix we describe the calibration of ACB-73, and we present data supporting the accuracy of the score scales for ACB-73 and ASVAB 2.

ACB-73 was used to select and classify Army recruits from mid-1973 until 1 January 1976, when ASVAB 5/6/7 was implemented. Because ACB-73 was used as a reference test for calibrating ASVAB 5/6/7, the accuracy of the ACB-73 score scale is critical for determining the source of the ASVAB miscalibration. The sample used to scale ACB-73 was 3,731 Army accessions. The sample was truncated because persons with low AFQT 7/8 scores, the reference test for calibrating ACB-73, had been rejected before entering the Army, and thus did not become accessions. AFQT 7/8 scores were missing for about one-third of the sample; they were estimated from the then operational form of the ACB (ACB-61). AFQT 7/8 was administered at AFEES before accessioning and ACB-61 at reception centers after accessioning. A statistical technique was used to estimate or fill in the bottom end of the score distributions. The ACB-73 score scale was developed on this filled-in sample.

The accuracy of the ACB-73 score scale was verified two ways in the sample used to calibrate ACB-73:

- The original General Technical (GT) composite scale from AFQT 7/8 agreed closely with the scale for the same composite in the ACB-73 calibration sample.
- The GT composite from ACB-73, referenced to AFQT 7/8 for developing the ACB-73 operational score scale, agreed closely with the GT scale from ACB-61.

The accuracy of the fill-in estimation technique was checked by applying it to a full-range sample that was truncated the same way as the ACB-73 calibration sample; the score scales from the full-range and the truncated sample from the two scales were in close agreement.

We also compared the score distributions of Army and Marine Corps applicants for enlistment obtained from the ACB and ASVAB, respectively, with those obtained from forms 7 and 8 of the Armed Forces Qualification Test.

The evidence indicates that the reference test scales (ACB-73 and ASVAB 2) were reasonably close to the traditional AFQT scale.

BACKGROUND

The Army Classification Battery (ACB) was used to classify recruits to skill specialties and to supplement the Armed Forces Qualification Test (AFQT) as an enlistment screen. The ACB was first used in 1949. The battery was revised periodically, based on new validation results for predicting performance in skill training courses. The last version of the ACB was implemented in 1973, and was called ACB-73. It remained in use until January 1976, when it was replaced by the Armed Services Vocational Aptitude Battery (ASVAB). Since 1976, the ASVAB has been used by all services for screening and classifying enlisted personnel.

Forms 2 and 3 of the ASVAB (ASVAB 2/3) were introduced in January 1973 and used until 1976, when they were replaced by forms 5, 6, and 7 (ASVAB 5/6/7). Form 2 was used in the high school testing program, and form 3 was used by the Marine Corps and Air Force to select and classify recruits. The content of forms 2 and 3 were parallel; each form had the same score scale [D-1].

Until 1973-74, a common AFQT was administered to all applicants for enlistment and registrants for induction. Since 1974, all services have obtained AFQT scores from their classification batteries. The Army stopped administering a separate AFQT in mid-1973. The Marine Corps stopped using a separate AFQT in 1974, when they started using ASVAB 3. From 1973 through 1975, all services did not use a common AFQT or classification test. With the introduction of ASVAB as the interservice test battery to screen and classify applicants, the use of a common AFQT score was restored. Contrary to earlier practice, however, the AFQT was embedded in the classification battery, and the subtests composing the AFQT (Word Knowledge, Arithmetic Reasoning, and Space Perception) were also used to help define aptitude composites.

In the early 1970s, it was decided to change the composition of AFQT. From 1953 until about 1973, the AFQT contained four item types: word knowledge, arithmetic reasoning, space perception, and tool knowledge. In the early 1970's, the tool knowledge items were dropped, and the three remaining item types were used to define the AFQT in ACB-73, ASVAB 2/3, and ASVAB 5/6/7.

The accuracy of the ACB-73 and ASVAB 2/3 score scales is critical. They were used as a measure of the quality of the All Volunteer Force (AVF) during the initial years of AVF, and they provide baseline data to help evaluate the operational effect of the miscalibration of ASVAB 5/6/7. The former consideration, quality of enlisted accessions as measured by the AFQT scores derived from ACB-73 and ASVAB 2/3, was used to show that the AVF was successful. The latter consideration, baseline data for evaluating the effects of ASVAB 5/6/7 miscalibration,

helps determine when the miscalibration occurred and the magnitude of the effect. If they were miscalibrated, then the score scale was in error during the early years of AVF, and ASVAB 5/6/7 may simply have continued the error. There is, of course, no miscalibration before 1973 because the AFQT then in use (forms 7 and 8) is the reference test for calibrating ACB-73, ASVAB 2/3, and ASVAB 5/6/7.

All other things equal, any miscalibration of a new test should be apparent by abrupt shifts in the AFQT score distribution when the new test is introduced. However, a simultaneous occurrence with the introduction of a new test is the temporary reduction in coaching or test compromise, which can mask the effects of miscalibration.

PROCEDURES FOR CALIBRATING ACB-73

When ACB-73 was developed and calibrated, it was designed only to classify Army recruits to skill specialties. The subtests in the battery were evaluated on the basis of their validity for predicting success in Army skill training courses [D-2]. The calibration sample was Army recruits tested in March and April 1972 at four Army reception centers.

While the calibration effort was underway, the Department of Defense decided that a separate AFQT was no longer required, and each service was permitted to derive an AFQT score from its classification battery. The AFQT score derived from ACB-73 was also calibrated on the sample of Army recruits tested at reception centers. The sample of recruits consisted of both inductees and enlistees. The draft was being phased out during the first half of calendar year 1972, and probably most of the sample was enlistees.

The experimental battery (ACB-73) was administered before the operational ACB (ACB-61). In this way, practice or fatigue effects that might result from taking ACB-73 after the operational tests were avoided. The ACB-73 scores, therefore, should be similar to those obtained under operational conditions. The tests were administered at reception centers rather than at AFEES. The recruits had already taken the AFQT at AFEES; those who failed to qualify for service had been removed and, of course, were unavailable for testing at reception centers.

In the 1972 data collection effort, the operational AFQT and ACB-61 scores were obtained from personnel records. The operational AFQT score distribution was examined to determine the score range in which there were sufficient cases to provide reliable frequency estimates. A statistical technique was developed to fill in the missing portion of the AFQT and ACB-73 score distributions. The fill-in technique estimates the frequency of scores for those portions of distributions that were truncated by prior selection of recruits on the AFQT. The fill-in technique, then, permits constructing the score scale for the entire

range of ability including those who fail to meet the selection standards. The technique relied on the correlation between each aptitude composite from ACB-73 (including the AFQT score) and the operational AFQT score (form 7 or 8); it assumed a normal bivariate distribution of AFQT and ACB-73 scores. The fill-in technique is described in detail elsewhere [D-3]; the procedure briefly is as follows:

- Compute the correlation between the operational AFQT score (form 7 or 8) and ACB-73 composite score; compute the standard error of estimate (SEE) for standard scores (SEE = $1 - r^2$) for predicting AFQT from aptitude composites. The correlation coefficient is equal to the slope of the regression line.

- Divide the distribution of ACB-73 composite scores into segments 0.2 standard deviations wide.

- Determine the portion of the distribution that failed the AFQT at the AFEES.

- Compute the distance, in normal deviate units, between the regression line at each segment of the ACB-73 scores and the AFQT passing score. Determine the percent in each segment of the ACB-73 composite score that passed the AFQT.

- Compute the cumulative percentage for the composite score, which has been truncated because the sample was selected on the AFQT.

- Convert the cumulative percentage for the composite score in the truncated sample (with AFQT failures removed) to the corresponding cumulative distribution in the full range sample (based on cumulative normal distribution). In this step, percentile scores in the truncated sample are equated to percentile scores in a full range sample.

The percentile scores computed from the last step are the correct values where the distribution of ACB-73 scores has been incidentally truncated because of the explicit truncation of AFQT 7/8 scores. This technique was applied to each of the aptitude composites and to the AFQT score derived from ACB-73. Each operational ACB-73 score was computed using this technique.

When the AFQT 7/8 scores were obtained from personnel records in 1972, many of the scores were missing. A multiple regression equation was developed to predict AFQT scores from ACB-61 subtest scores. Because the content of AFQT 7/8 was so similar to that of ACB-61, the correlation was high, and the prediction was accurate.

For purposes of this report, we made two checks on the accuracy of calibrating ACB-73. One check was to compare the calibration of the Word Knowledge and Arithmetic Reasoning subtests from ACB-73 with similar subtests from AFQT 7/8 and ACB-61. The second was to determine if the fill-in technique in a truncated sample reproduced the scale in a full-range sample: (a) truncate the sample on AFQT, (b) apply the statistical technique to fill in the missing portion of an aptitude composite; and (c) compare the calibration of the composite in the full-range and filled-in samples.

RESULTS

ACB-73 was administered to Army accessions at four Army reception centers, located in New Jersey, South Carolina, Missouri, and California. Of the total, 2,108 were administered form A of ACB-73 and 1,698 form B. Following administration of ACB-73, all of the sample also took the then operational form of the ACB (ACB-61). The AFQT scores (forms 7 and 8) of record were obtained from personnel records for all recruits who had their scores recorded. AFQT 7/8 scores were missing for 1,258 cases.

A regression equation was computed to predict AFQT 7/8 scores from ACB-61 subtests. The multiple correlation was .93. The distribution of AFQT scores is shown in table D-1 for cases with predicted scores and those with AFQT scores of record.

When the operational ACB-73 scores were developed, the two sets of AFQT scores were combined, and the combined AFQT distribution was used as the reference variable for calibrating ACB-73. The distribution of AFQT scores for the combined sample is also shown in table D-1. Examination of the AFQT score distribution showed that the number of cases below the 16th percentile score was too small to provide reliable estimates of the frequencies at each score. Explicit selection on AFQT 7/8, therefore, was assumed to occur below the 16th percentile score; only cases with an AFQT score of 16 and above, either observed or predicted, were retained in the sample.

This truncated sample was used in the analysis to develop the ACB-73 score scale for operational use. The sample was weighted to obtain a uniform distribution of AFQT scores. These weights, sometimes called stratification weights, are shown in the final column of table D-1. The fill-in technique, described above in the procedures section, was applied to the sample for calibrating each aptitude composite and AFQT 73.

The calibration of two ACB-73 composites is described in detail: one is the AFQT score obtained from ACB-73, called AFQT 73, because AFQT 73 scores were used as the reference for calibrating ASVAB 5/6/7; the other is the General Technical (GT) score, because independent checks on the calibration of GT can be accomplished. Both AFQT 73 and

GT contain the Word Knowledge (WK) and Arithmetic Reasoning (AR) subtests; AFQT 73 also contains the Space Perception (SP) subtest. Each of these subtests has 20 items, which means that the GT composite has 40 items and AFQT 73, 60 items.

TABLE D-1

DISTRIBUTION AND WEIGHTS OF AFQT 7/8 REFERENCE TEST

AFQT decile	Percent in decile: Observed[a]	Predicted[b]	Combined	Weight[c]
1-9	0	0	0	0
10-19	4	2	4	1.33
20-29	9	7	8	1.29
30-39	14	12	13	0.78
40-49	12	13	13	0.85
50-59	13	16	14	0.73
60-69	13	19	15	0.72
70-79	12	15	13	0.82
80-89	14	10	12	0.84
90-99	9	5	8	1.36
N	2,548	1,258	3,806	

[a]AFQT 7/8 scores obtained from personnel records.
[b]AFQT 7/8 scores predicted from ACB-61.
[c]Weights of AFQT 7/8 scores to obtain uniform score distribution; weight of 1.33 in decile 10-19 applied only to scores 16-19.

Checking the Calibration of the GT Composite

The first check we made of the GT composite was to compare the calibration of the Word Knowledge and Arithmetic Reasoning items from AFQT 7/8 in the ACB-73 calibration sample with the original calibration of AFQT 7/8, completed in 1959 [D-4]. The four subtests of AFQT 7/8 (Word Knowlege, Arithmetic Reasoning, Space Perception, and Tool Knowledge) had been calibrated separately in the original 1959 sample because they were also part of the Army Qualification Battery, used from the early 1960s until July 1973 to help determine the qualification of persons in AFQT category IV (percentile scores 10 through 30). Two conversions of the GT score from AFQT 7/8 were available: one computed in 1972 from the ACB-73 truncated sample of Army recruits and the second from the original 1959 calibration of AFQT 7/8. The two conversions are shown in figure D-1.

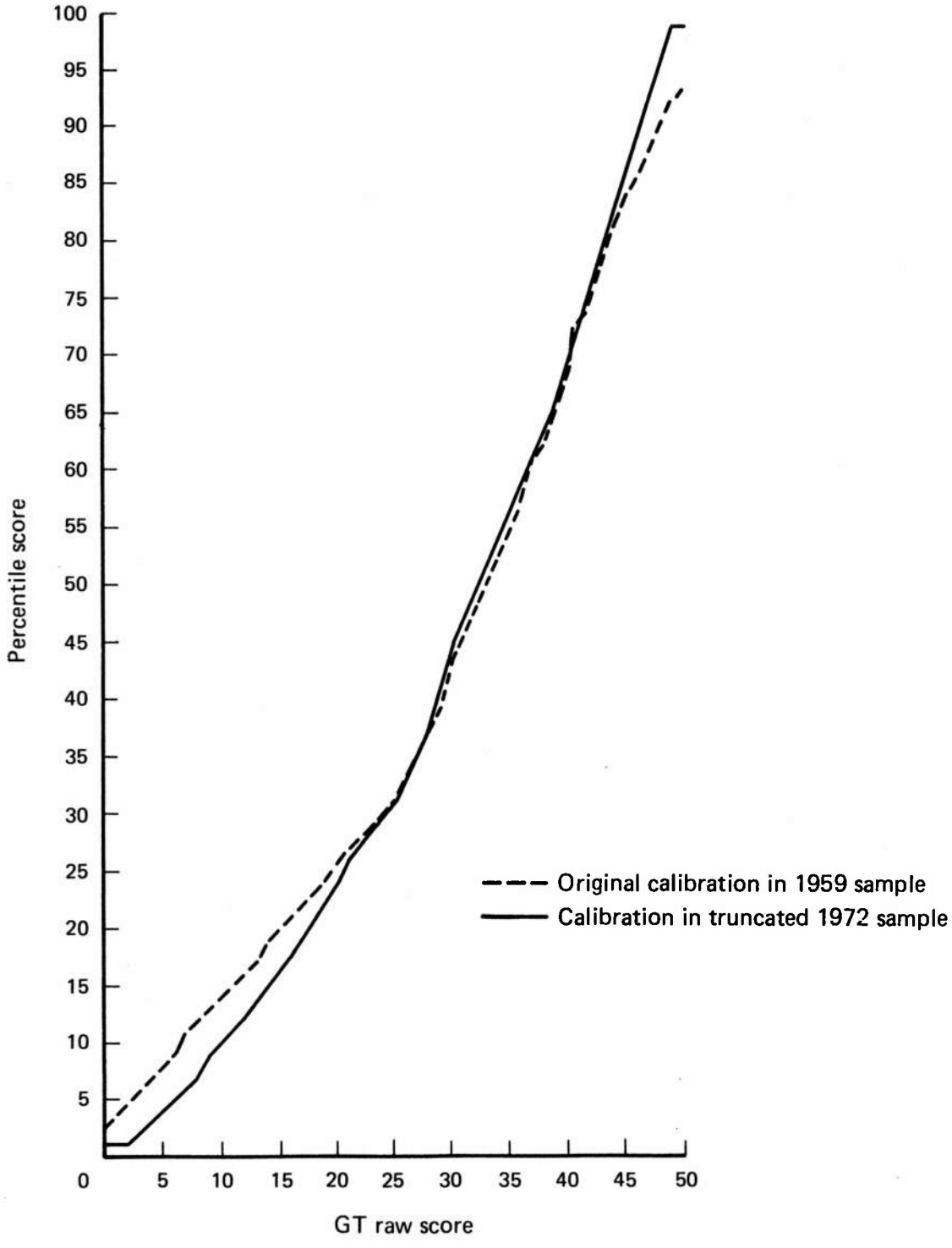

FIG. D-1: CALIBRATION OF GT SCORE FROM AFQT 7/8

The AFQT 7/8 GT conversion from the truncated ACB-73 calibration sample agrees closely with the original AFQT 7/8 calibration. The calibration in the ACB-73 sample is somewhat lower than the original in the bottom of the score range, up to a percentile score of 28, but then the conversions are similar until the upper end of the scale, above a percentile score of 75. This comparison supports the accuracy of the ACB-73 score scale because the original calibration of AFQT 7/8 was essentially reproduced in the ACB-73 sample of Army recruits. Because all ACB-73 composites, including AFQT, were calibrated in the same way as the GT score from AFQT 7/8, the correctness of the GT scale supports the correctness of other scales for ACB-73.

The second check we made for purposes of this report was to compare the operational conversion table for GT from ACB-73 to that of the counterpart subtests, called Verbal and Arithmetic Reasoning, in ACB-61. The use of ACB-61 subtests as reference variables is justified because ACB-61 was intended to be on the same scale as AFQT 7/8. The calibration of the ACB-73 GT composite using the Verbal and Arithmetic Reasoning subtests from ACB-61 as the reference variable is shown in table D-2. Each of the subtests in the ACB-61 GT composite, Verbal and Arithmetic Reasoning, was used as a separate reference variable. The cumulative frequencies for these variables are shown in table D-3. As shown in table D-2, the score scale using the Verbal test as the reference variable is higher than the scale using Arithmetic Reasoning as the reference. The column labeled "ACB-61 GT" is the mean of the Verbal and Arithmetic Reasoning columns. The score scales in table D-2 are expressed as standard scores. We converted the standard scores to percentile scores, using the traditional reference population. We plotted the conversion from GT raw score to percentile score for the operational ACB-73 GT scale and for the ACB-73 GT scale calibrated to GT from ACB-61; the conversions are shown in figure D-2. The two calibrations of ACB-73 GT, the operational one completed in 1972 and the one completed for this report using ACB-61 GT as the reference variable, are in general agreement.

A disturbing result from the using the ACB-61 subtests as the reference variables is that the scales for Verbal and Arithmetic Reasoning do not agree with each other. If both tests were on the same scale, then the conversion from ACB-73 GT raw score to the standard score for both subtests would be identical. But the conversion using Verbal as the reference test is systematically higher than when Arithmetic Reasoning is the reference test. This fact means that in representative samples, the mean score of the Verbal subtest would be on the order of six standard score points higher than the mean of the Arithmetic Reasoning subtests. These results show that the subtests in ACB-61 were not on a common metric, and therefore, any subtest may or may not

TABLE D-2

CALIBRATION OF ACB-73 GENERAL TECHNICAL (GT) COMPOSITE, USING ACB-61 GT COMPOSITE AS REFERENCE VARIABLE

	Standard score[a]			
		ACB-61		
ACB-73 GT raw score	ACB-73[b] GT	Verbal[c] test	Arithmetic[c] test	ACB-61[c] GT
0-10	62	62		
11	66	69	50	60
12	70	72	65	68
13	75	75	69	72
14	80	79	73	76
15	84	82	76	79
16	87	86	79	82
17	90	89	82	85
18	92	92	85	88
19	94	96	89	92
20	96	99	92	95
21	98	101	95	98
22	100	103	97	100
23	102	106	99	103
24	104	108	102	105
25	106	110	104	107
26	108	112	107	110
27	110	114	110	112
28	112	116	113	115
29	114	118	116	117
30	116	120	118	119
31	118	122	120	121
32	120	124	122	123
33	123	126	125	126
34	126	129	129	129
35	130	131	132	132
36	135	134	136	135
37	140	136	140	138
38	146	140	144	142
39	147	145	148	
40	151	150	152	

[a] Standard score scale has mean of 100 and standard deviation of 20.
[b] Operational ACB-73 GT standard score scale, computed in truncated sample.
[c] GT composite in ACB-61 is mean of Verbal and Arithmetic Reasoning tests. Separate calibration computed for each ACB-61 test and then averaged to obtain GT conversion.

TABLE D-3

CUMULATIVE FREQUENCIES OF SCORES
FOR THE GENERAL TECHNICAL (GT) COMPOSITE

ACB-73 GT composite		ACB-61 GT subtests					
			Cumulative percent			Cumulative percent	
Raw score	Cumulative percent	Standard score	Arithmetic reasoning	Verbal	Standard score	Arithmetic reasoning	Verbal
0-8	1	60-1	4	2	122-3	82	79
9	1	62-3	4	2	124-5	84	83
10	2	64-5	5	3	126-7	86	86
11	3	66-7	6	3	128-9	87	87
12	5	68-9	7	4	130-1	89	90
13	7	70-1	9	5	132-3	92	92
14	9	72-3	10	6	134-5	94	95
15	12	74-5	12	7	136-7	95	96
16	15	76-7	13	8	138-9	95	97
17	19	78-9	15	9	140-1	96	98
18	23	80-1	18	11	142-3	97	98
19	28	82-3	21	13	144-5	98	99
20	32	84-5	23	15	146-7	99	99
21	36	86-7	25	17	148-9	99	99
22	40	88-9	27	19	150	100	100
23	45	90-1	31	21			
24	50	92-3	34	24			
25	55	94-5	37	26			
26	59	96-7	41	29			
27	63	98-9	45	32			
28	67	100-1	49	36			
29	71	102-3	52	40			
30	75	104-5	55	43			
31	78	106-7	59	47			
32	81	108-9	62	51			
33	84	110-1	64	56			
34	88	112-3	67	60			
35	91	114-5	70	64			
36	94	116-7	72	68			
37	96	118-9	75	72			
38	98	120-1	79	76			
39	99						
40	100						

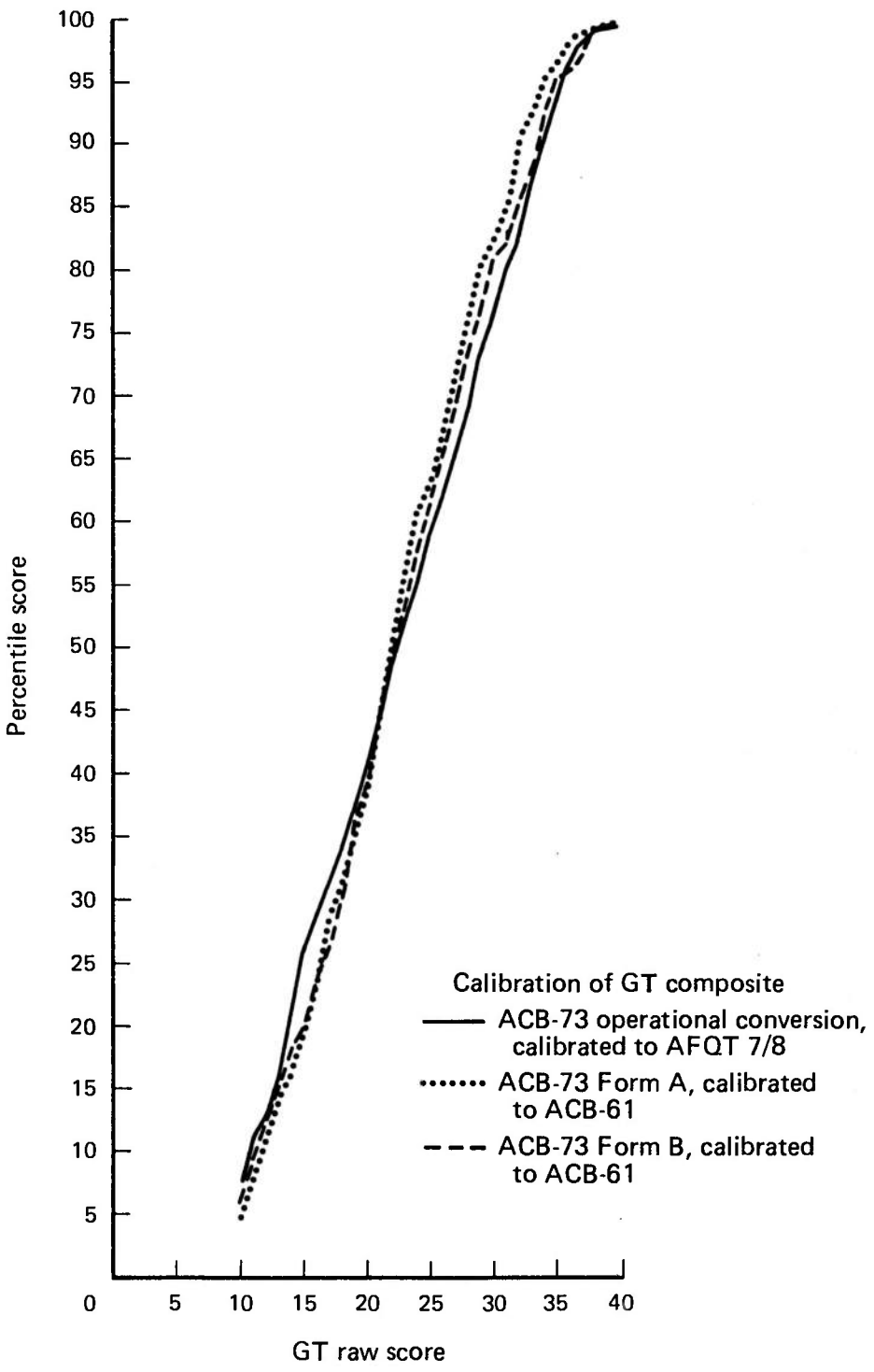

FIG. D-2: CALIBRATION OF GT COMPOSITE FROM ACB-73

be on the AFQT score scale.* Thus, no definitive conclusions can be drawn about the accuracy of the ACB-73 scale from the similarity, or difference, between the GT scales from ACB-61 and ACB-73.

Further data on the accuracy of the ACB-73 scale is contained in a report comparing AFQT 7/8 and ACB-73 scores in a sample of 305 Army recruits [D-5]. Based on this small sample tested in fall 1973, 35 percent was placed in categories IV and V by ACB-73 compared to 45 percent by AFQT 7/8.

Accuracy of the Fill-In Technique

We checked the accuracy of the fill-in technique by comparing the calibration of AFQT in a full-range sample of applicants for enlistment with the calibration in the same sample truncated by deleting the bottom of the AFQT score range. A sample of 5,069 applicants was tested in June and July 1979 with the then operational version of the ASVAB (forms 6 and 7) and with AFQT 7A. AFQT 7A was the reference test used to check the calibration of ASVAB 5/6/7 [D-6], to calibrate ACB-73, and to calibrate the new versions of ASVAB (forms 8, 9, and 10) implemented in October 1980.

The full-range sample was truncated by deleting all cases with AFQT 7A percentile scores below 21. We applied the fill-in technique to the truncated sample, and compared percentile scores for the full-range and truncated samples. The results are shown in figure D-3.

The two conversions for AFQT from ASVAB 5/6/7 were similar; but, contrary to the results for checking the accuracy of the fill-in technique on a truncated sample of Army recruits, there was no consistent inflation or depression of converted scores in AFQT categories IV and V. The conversion lines cross at percentile scores of 6 and 15. The conclusion, then, is that the fill-in technique did not produce any systematic bias in the calibration of ACB-73.

SCORE DISTRIBUTIONS OF ACB-73 AND ASVAB-3

By comparing score distributions of applicants for enlistment obtained from new tests (ACB-73 and ASVAB 3) with those from a test known to have an accurate scale (AFQT 7/8), we can gain additional information about the accuracy of the scale for the new tests.

---

* Each subtest in the ACB-61 battery was calibrated separately. Also the subtests were calibrated on samples of Army recruits rather than on full-range samples of applicants. All subtests in each form of the ASVAB, by contrast, have always been calibrated at the same time. All ASVAB subtests in each form therefore are on the same metric. Similarly, all subtests in ACB-73 were calibrated at the same time.

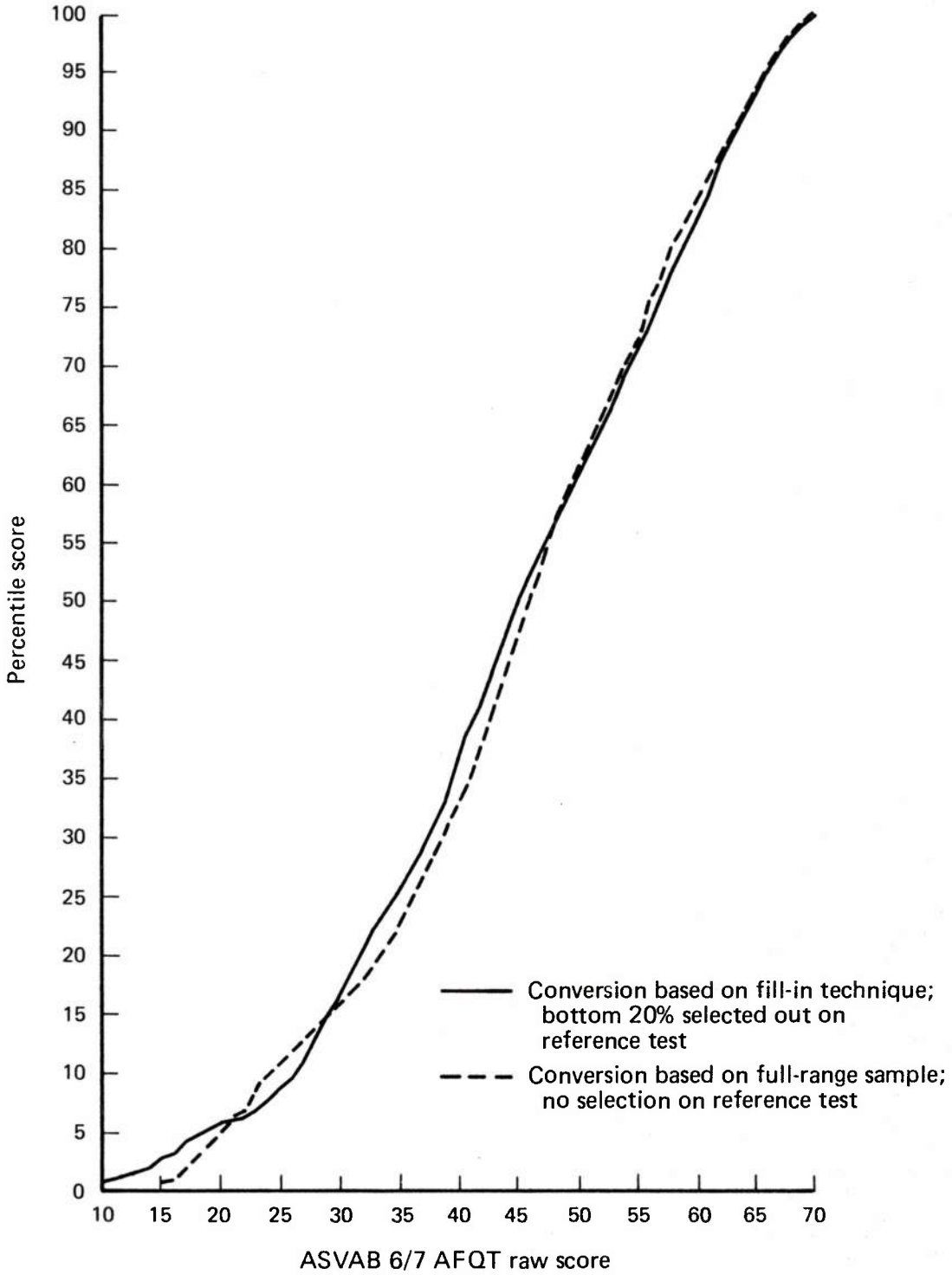

FIG. D-3: CALIBRATION OF AFQT FROM ASVAB 6/7 IN TRUNCATED AND FULL RANGE SAMPLE OF 1979 APPLICANTS

We obtained AFQT score distributions of Army applicants for the last 3 months that AFQT 7/8 was used and the first 3 months that ACB-73 was used to obtain AFQT scores.* ACB-73 replaced AFQT 7/8 in the Army on 1 July 1973.

We also obtained the AFQT score distributions for Marine Corps applicants for the last 3 months when AFQT 7/8 was used and the first 3 months when ASVAB 3 was used. The Marine Corps replaced AFQT 7/8 with ASVAB-3 on 1 July 1974. The accuracy of the scale of ASVAB 3 is relevant because it is parallel to ASVAB 2, the reference test used for the sample of Navy and Air Force recruits. The AFQT score distributions for the Army and Marine Corps applicants are shown in figure D-4. The pattern of changes in AFQT score distributions for the two tests used by the two services is remarkably similar. In both cases, the percentage of applicants placed in categories IVB, IVC, and V increased when the new test was introduced. These increases are expected when a compromised test (AFQT 7/8) is replaced by a new test. The percentage of applicants placed in category IVA decreased for ACB-73 and remained relatively stable for ASVAB 3; these changes could be a function of test compromise or inflation of the scale in category IVA, or some combination.

In category IIIB, we observed a sudden increase of about 4 percentage points for both tests, from about 20 to 24 percent for ACB-73 and 21 to 25 percent for ASVAB 3. Given that AFQT 7/8 was compromised, this increase was unexpected. Typically, test compromise moves the bulk of those who rightfully belong in categories IV and V into category IIIB, and when a new test is introduced, the percentage in category IIIB should drop. If the new tests were on the same scale as AFQT 7/8, we would have expected the percentage in category IIIB to decline. The increase of examinees in category IIIB indicates that either test compromise did not move examinees into category IIIB or the scales for ACB-73 and ASVAB 3 were in error by placing an excessive number in category IIIB. Based on our analysis of coaching presented in the main text, the most likely explanation of the increase is that the scales for ACB-73 and ASVAB 2/3 placed too few examinees in category IVA and too many in IIIB. In other words, the problem seems to be that the scales are inflated in category IVA and too difficult in category IIIB.

* The Army used ACB-73 in May 1973 to obtain aptitude composite scores. Both ACB-73 and AFQT 7/8 were administered to all Army applicants during May and June 1973. In July 1973, the Army suspended using AFQT 7/8.

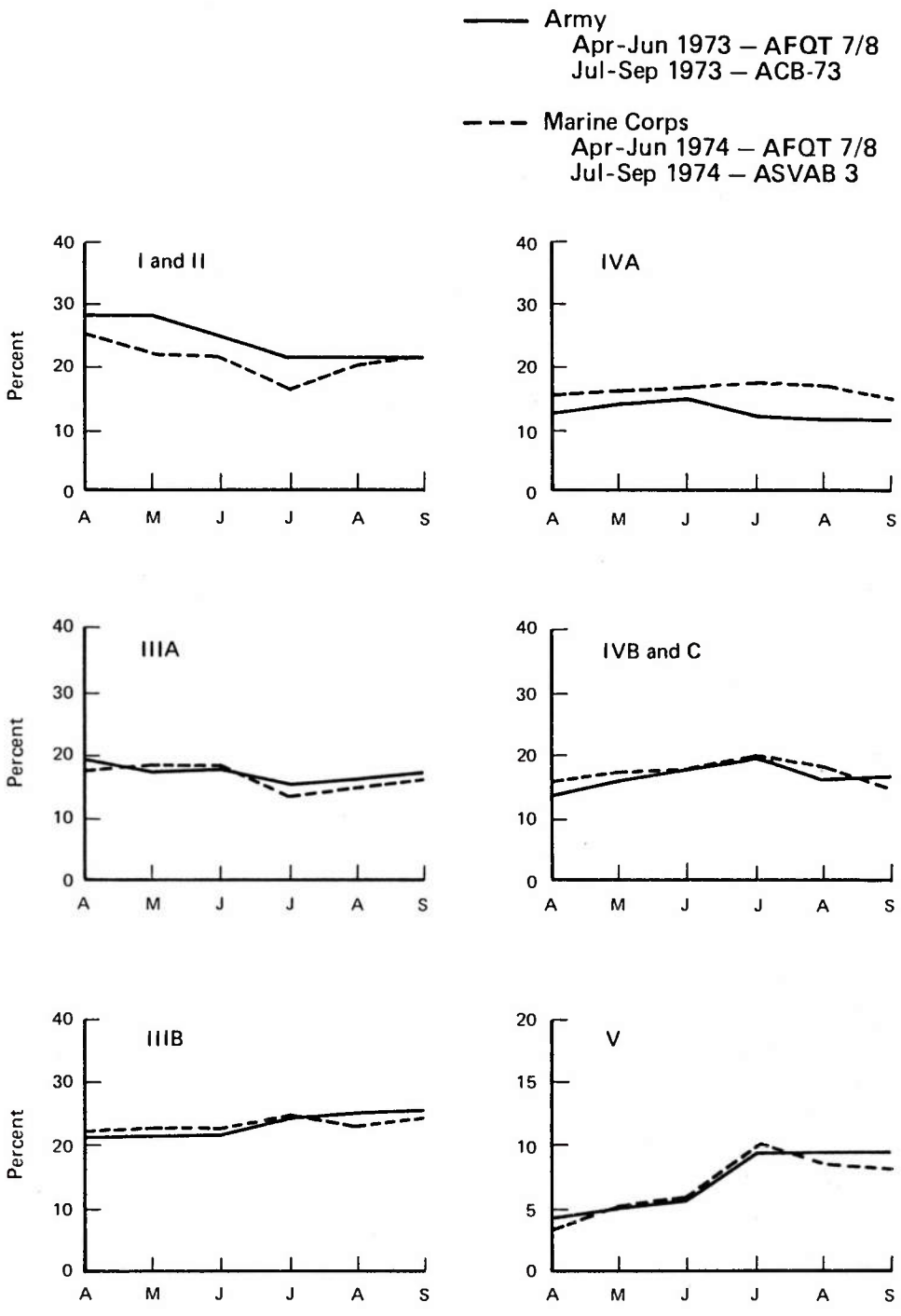

FIG. D-4: SCORE DISTRIBUTIONS OF ARMY AND MARINE CORPS APPLICANTS AT TIME OF TRANSITION FROM AFQT 7/8

DISCUSSION

The general accuracy of the ACB-73 score scale is also supported by the operational experience during the 2.5 years the battery was used to select and classify Army recruits--from July 1973 until January 1976. Immediately after ACB-73 was implemented, the Army Training Command (then called Continental Army Command) reduced aptitude composite prerequisites for many of the skill training courses. Until that time, the bulk of the courses had a prerequisite score of 100 on the Army standard score scale, which corresponds to a percentile score of 50; most of these prerequisites were lowered to a standard score of 90, which corresponds to a percentile score of 31. While ACB-73 was in use, there was no large scale increase in the prerequisites. The combination of lowered aptitude composite prerequisites plus generally acceptable failure rates in the training courses suggests that ACB-73 score scale was functioning as expected by the Army and that the ACB-73 score scale was not seriously inflated in the region around the minimum prerequisite scores.

The experience of Army trainers was opposite after ASVAB 5/6/7 was implemented. During 1977 and 1978, the Army increased the prerequisites, usually from 90 to 100, for more than 60 skill specialties because the failure rates were excessive.

The results suggest that, in general, ACB-73 and ASVAB 2/3 were calibrated correctly. There is some evidence, however, that their scales are inflated in categories IVA and IIIB. Because of uncertainties in the data, we cannot estimate accurately how large the inflation is. We can be sure, though, that the inflation of these two tests, used as the reference tests for calibrating ASVAB 5/6/7, is not enough to explain the inflation of the original ASVAB 5/6/7 scale.

REFERENCES

[D-1] Army Research Institute, Technical Paper 289, "Development of the Armed Services Vocational Aptitude Battery (ASVAB) Forms 2 and 3," by Leonard C. Seeley, M. A. Fischl, and Jack M. Hicks, Unclassified, Feb 1978

[D-2] Army Research Institute, Technical Research Note 239, "The Development and Evaluation of a New Army Classification Battery," by Milton H. Maier and E. F. Fuchs, Unclassified, Sep 1972

[D-3] Army Research Institute, Technical Research Report 1109, "Standardization of Common Core Tests," by James B. Trump et al., Unclassified, Dec 1957

[D-4] Army Research Institute, Technical Research Report 1132, "Development of the Armed Forces Qualification Tests 7 and 8," by A. G. Bayroff and A. A. Anderson, Unclassified, May 1963

[D-5] Army Research Institute, Research Memorandum 74-5, "A Comparison of Three-Subtest AFQT and Four-Subtest AFQT," by Milton H. Maier, Unclassified, Mar 1974

[D-6] Office of the Seretary of Defense, (Directorate for Accession Policy) Technical Memorandum 80-1, "Renorming ASVAB 6/7 at Armed Forces Examining and Entrance Stations," by Milton H. Maier and Frances C. Grafton, Unclassified, Aug 1980

APPENDIX E

CALIBRATION OF ASVAB 5/6/7 FOR ARMY EXAMINEES GROUPED BY DATE OF TESTING WITH ACB-73

APPENDIX E

CALIBRATION OF ASVAB 5/6/7 FOR ARMY EXAMINEES GROUPED BY DATE OF TESTING WITH ACB-73

The reference test (ACB-73) score distribution in the original calibration sample of Army examinees indicated that many applicants who failed to qualify for enlistment were not tested with the experimental ASVAB 5/6/7. One way of verifying such selection is to determine the date that ACB-73 scores were entered into the examinees' records. The Defense Manpower Data Center (DMDC) provided date of testing and AFQT scores of record for 1,880 of the 2,512 Army male examinees in the original sample. The scores of these 1,880 cases were used to estimate the extent of prior selection on the reference test and the effects of selection on the score scale.

No testing with ASVAB 5/6/7 took place before 1 September 1975 and the testing was virtually completed by 31 October 1975 [E-1]. At some AFEES the testing may have started after 15 September 1975. The Army sample was divided into four groups based on the date of the ACB-73 testing.

Group 1 - Scores of record dated 1 January-31 August 1975; this group was selected on the basis of ACB-73 scores; therefore few ACB-73 percentile scores should be below 16 (minimum qualifying score for enlistment) or above 50 (maximum score that counted toward meeting the AFEES quotas).

Group 2 - Scores of record dated 1-15 September 1975; expect more low (below 16) and high (above 50) scores than in group 1.

Group 3 - Scores of record dated 16 September-31 October 1975; expect least selection on basis of ACB-73 scores in this group, and therefore the most low and high ACB-73 scores.

Group 4 - Unknown date for scores of record or dates after 31 October 1975; no expectations about distribution of scores in this group.

The mean, standard deviation, and correlation between the AFQT scores of record and scores coded on the answer sheets are shown in table E-1. The means and standard deviations in all groups were about equal. The correlation between the two sets of scores was high (coefficients greater than .9) in groups 2, 3, and 4. In group 1, the correlation was lower (coefficient of .77).

TABLE E-1

ACCURACY OF CODING ACB-73 SCORES ON
ASVAB 5/6/7 ANSWER SHEETS

Date of ACB-73 score	Mean	Standard deviation	Correlation coefficient[a]
Group 1 (score of record dated 1 Jan - 31 Aug 1975), N = 556			
Score of record	38.3	10.9	.77
Coded score	38.7	10.5	
Group 2 (score of record dated 1-15 Sept 1975), N = 519			
Score of record	36.4	20.5	.93
Coded score	35.7	19.9	
Group 3 (score of record dated 16 Sep - 31 Oct 1975), N = 632			
Score of record	38.8	20.8	.94
Coded score	37.7	20.4	
Group 4 (unknown or later date), N = 173			
Score of record	36.9	18.7	.95
Coded score	36.2	18.5	

[a]Correlation between AFQT scores coded on answer sheets and scores of record.

The correlation coefficients between the AFQT 73 scores of record and AFQT scores from ASVAB 5/6/7 in groups 1 through 4 were .40, .73, .75, and .70, respectively. The correlation coefficients between the AFQT scores coded on the answer sheet and the AFQT scores from ASVAB 5/6/7 in groups 1 through 4 were .33, .69, .75, and .62, respectively.

The cumulative frequencies of the ACB-73 percentile scores and the ASVAB 5/6/7 raw scores for the total Army sample are shown in table E-2. The number of matched scores with AFQT scores in groups 1 through 4 is 556, 519, 632, and 173, respectively. These distributions show that group 1 did have relatively few persons with ACB-73 scores in categories IV and V (percentile scores 1-30) or above 50. Groups 2 and

TABLE E-2

CUMULATIVE FREQUENCIES OF ACB-73 AND ASVAB 5/6/7 SCORES

	ACB-73 percentile score						ASVAB 5/6/7 raw score					
	Cumulative frequency						Cumulative frequency					
Percentile score	Total[a]	1[b]	2[c]	3[d]	4[e]		Raw score	Total[a]	1[b]	2[c]	3[d]	4[e]
0-5	1	0	2	1	1		0-12	1	1	1	1	1
6	3	1	4	2	2		13	1	1	1	1	1
7	3	1	5	3	3		14	2	1	2	1	1
8	4	1	6	4	4		15	2	2	3	2	1
9	6	1	8	6	6		16	3	2	4	3	1
10	8	1	10	8	7		17	4	2	5	3	2
11	8	1	10	8	8		18	6	3	6	4	4
12	10	2	13	11	10		19	8	6	9	5	6
13	10	2	14	11	10		20	10	7	12	7	8
14	13	3	16	13	12		21	12	8	14	9	10
15	13	3	16	13	12		22	14	9	16	11	13
16	16	5	20	15	13		23	16	11	18	13	16
17	16	5	20	15	13		24	19	13	21	18	21
18	18	6	23	18	17		25	23	15	24	21	24
19	18	6	23	18	17		26	26	17	28	24	28
20	18	6	23	18	17		27	30	20	34	28	32
21	22	9	28	22	24		28	33	23	38	32	36
22	23	9	29	22	24		29	36	26	41	36	41
23	23	9	29	23	24		30	40	30	45	40	43
24	23	9	30	23	24		31	44	34	48	44	46
25	27	11	32	28	27		32	48	39	53	48	52
26	28	11	33	29	28		33	52	44	58	52	54
27	29	12	33	29	29		34	55	49	60	55	58
28	31	15	37	32	31		35	58	52	62	58	64
29	31	15	37	32	31		36	62	55	65	61	68
30	32	17	38	32	31		37	65	60	68	64	71
31	39	25	43	39	41		38	68	65	70	67	73
32	39	26	44	39	41		39	70	69	72	70	75
33	44	34	49	45	47		40	73	72	75	72	78
34	44	34	50	45	48		41	76	76	77	75	79
35	46	37	50	47	48		42	79	80	78	77	82
36	51	43	54	54	54		43	81	82	82	80	84
37	52	45	55	54	56		44	83	85	83	81	87
38	57	49	59	60	64		45	84	88	83	82	88
39	59	54	62	61	64		46	86	89	86	83	89

TABLE E-2 (Cont'd)

	ACB-73 percentile score					ASVAB 5/6/7 raw score					
	Cumulative frequency					Cumulative frequency					
Percentile score	Total[a]	1[b]	2[c]	3[d]	4[e]	Raw score	Total[a]	1[b]	2[c]	3[d]	4[e]
40	60	54	62	61	64	47	88	91	88	85	90
41	64	59	66	65	69	48	89	93	89	87	92
42	65	62	67	66	70	49	91	94	90	88	94
43	65	63	67	66	70	50	92	96	91	89	94
44	72	72	72	71	75	51	93	97	92	90	94
45	72	73	72	71	75	52	94	98	92	91	95
46	73	75	72	72	75	53	94	98	93	92	95
47	78	82	77	75	80	54	95	98	93	93	96
48	80	86	78	75	82	55-70	100	100	100	100	100
49	81	87	78	75	82						
50	85	96	82	77	83						
51	85	96	82	78	83						
52	85	96	82	78	85						
53	86	97	83	79	86						
54	86	97	83	79	86						
55	87	97	84	80	86						
56	87	97	85	81	87						
57	88	97	86	81	87						
58	88	98	86	82	87						
59	89	98	87	83	88						
60	89	98	88	83	88						
61	90	98	89	84	88						
62	90	98	89	85	88						
63	91	99	90	86	89						
64	91	99	90	87	90						
65	92		91	88	91						
66	92		91	89	91						
67	92		91	89	92						
68	93		92	90	93						
69	93		92	90	93						
70	94		92	90	93						

[a] N = 2,512; ACB-73 scores coded on ASVAB answer sheets.
[b] N = 556; tested 1 January through 31 August 1975 with ACB-73.
[c] N = 519; tested 1 through 15 September 1975 with ACB-73.
[d] N = 632; tested 16 September through 31 October 1975 with ACB-73.
[e] N = 173; tested after October 1975 with ACB-73.

3 had more persons with low or high scores; but, contrary to expectations, they were similar to each other.

In groups 1, 2, and 3, the AFQT scores from ASVAB 5/6/7 were scaled to the ACB-73 scores of record using the equipercentile equating technique. The scales are shown in table E-3. As expected, the scale for group 1 shows the most inflation. The inflation was pronounced in category IV (percentile scores 10-30). In the category IV range, the scales for groups 2 and 3 were similar to that for the total Army sample, shown in the first column of table E-3. The scale for the total Army sample is inflated, compared to the correct scale (shown in the last column of table E-3) until about a percentile score of 50. The scales for the groups 1, 2, and 3 merged near a percentile score of 30. At about the 50th percentile score, the scale for group 1 converged with correct scale (raw score 46 = percentile score 50), but the inflation for groups 2 and 3 persisted.

The calibration for group 4 used the AFQT 73 score coded on the answer sheet as the reference test; the resulting scale was similar to that for groups 2 and 3. The calibration for groups 1, 2, and 3 were similar whether the AFQT-73 scores of record or the AFQT-73 scores coded on the answer sheet were used as the reference variable.

These results suggest that prior selection on the reference test produced some inflation in the score scale. But, the fact that the scale for group 1 converged with the scale for the other groups and that groups 2 and 3 were similar to the total Army sample suggests that selection by itself does not account for all of the inflation. The maximum difference between the correct scale and the scale based on groups 2 and 3 is 16 to 18 percentile score points, in the raw score range of 30 through 35, and percentile score 17 through 23 range in the correct scale. The scale for the Army sample is inflated, whether based on the total Army sample or just on groups 2 and 3.

TABLE E-3

CALIBRATION OF ASVAB 5/6/7 IN ARMY SAMPLE, GROUPED BY DATE OF TESTING WITH ACB-73 COMPARED TO CALIBRATION IN FULL SAMPLE AND CORRECT SCALE

ASVAB raw score	Percentile score					
	Full Army sample[a]	Group 1[b]	Group 2[b]	Group 3[b]	Group 4[c]	Correct
0-13	3	4	5			
14	5	5	5			
15	6	10	5.5	6	6	
16	7	11	6	6.5	7	
17	7	12	7	7	8	2
18	8	14	8	7	9	3
19	10	17	10	8	10	4
20	11	18	11	9	11	5
21	13	20	12.5	11	13	6
22	15	22	14	12	15	7
23	17	25	15	14	17	9
24	19	26	17	18	20	10
25	21	28	19	20	23	11
26	24	29	21	22	26	12
27	27	29	26	25	28	13
28	29	30	29	27	30	14
29	30	31	30	30	31	15
30	32	32	31	31	32	16
31	33	33	33	32	35	17
32	35	35	35	34	37	18
33	36	36	37	35	38	19
34	37	38	39	37	40	21
35	38	39	40	38	41	23
36	40	40	41	39	43	25
37	41	41	42	41	45	27
38	42	43	43	42	46	29
39	43	44	44	43	47	31
40	45	45	45	45	47	33
41	46	46	46	48	48	35
42	47	46	48	50	49	38
43	48	47	50	54	51	41
44	49	48	52	56	53	44
45	50	49	54	58	56	47

TABLE E-3 (Cont'd)

ASVAB raw score	Percentile score: Full Army sample[a]	Group 1[b]	Group 2[b]	Group 3[b]	Group 4[c]	Correct
46	54	49	56	58	60	50
47	58	49	59	62	64	53
48	60	49	62	64	67	56
49	60	49	62	64	67	56
50	66	50	65	67	74	60

[a]ASVAB 5/6/7 conversion based on full Army sample, N=2,512, using equipercentile equating technique.
[b]ACB-73 score of record used as reference variable.
[c]ACB-73 score coded on answer sheet used as reference variable.

REFERENCE

[E-1] Office of the Assistant Secretary of Defense (Manpower, Reserve Affairs, and Logistics), "History of the Armed Services Vocational Aptitude Battery (ASVAB), 1974-1980," ASVAB Working Group, Unclassified, Mar 1980

APPENDIX F

EFFECTS OF SIMULATING SELECTION ON AN OPERATIONAL REFERENCE TEST

APPENDIX F

EFFECTS OF SIMULATING SELECTION ON AN OPERATIONAL REFERENCE TEST

Appendix E shows that prior selection of examinees based on their reference test scores produced an inflated score scale; because other factors operated in 1975 to inflate the score scale, the effects of selection on the operational reference test cannot be isolated and quantified in the original 1975 sample. The effects can, however, be estimated from the data collected on a sample of 1980 applicants used to calibrate ASVAB 8A.

ASVAB 8A, a replacement for ASVAB 5/6/7, was calibrated on a sample of applicants for all services. The experimental testing was conducted in January and February 1980. AFQT 7A was used as the reference test for computing the ASVAB 8A score scale. All experimental tests (ASVAB 8A and AFQT 7A) were administered before the operational test (ASVAB 6 or 7). The studies to calibrate ASVAB 8A are reported by Sims and Truss [F-1] and Maier and Grafton [F-2].

We also collected operational ASVAB 5/6/7 scores of the applicants who took the experimental ASVAB 8A and AFQT 7A. For the purpose of this analysis we used their operational ASVAB 5/6/7 scores as the reference test. Prior selection on a reference test was simulated by weighting the ASVAB 5/6/7 score distribution in the 1980 sample, which includes applicants for all services, to have the same distribution as the ACB-73 scores in the 1975 sample of Army examinees tested during September and October. Our assumption is that the weighted distribution of ASVAB 5/6/7 scores in the 1980 sample simulates the effects of selecting the Army examinees in the 1975 sample on the basis of their operational ACB-73 scores. The distribution of ASVAB 8A raw scores was computed in the weighted sample. In effect, the ASVAB 5/6/7 percentile scores were weighted explicitly to simulate the ACB-73 scores of the Army examinees tested in September and October 1975. The distribution of ASVAB 8A was affected by the weighting to a similar degree as the distribution of ASVAB 5/6/7 in the 1975 Army sample. We made a check to ensure that the exclusive use of Army applicants in the 1975 calibration sample did not produce an inflated score scale. We calibrated ASVAB 5/6/7 to AFQT 7A in a sample of Army applicants tested in 1979 and compared the scale to that obtained on the complete sample of applicants for all services. The results are identical.

The distributions of reference test scores in the 1975 and 1980 samples are shown in table F-1. As expected, the 1980 sample had more examinees with AFQT percentile scores below 20 and above 60. The weights for the 1980 sample, accordingly, are less than one for deciles in this range. In the range of percentile scores 30-49, the 1980 sample had significantly fewer cases; the weights in this range are much larger

than one. (The ASVAB 5/6/7 scores in the 1980 sample are based on the correct score scale.) The frequencies in each AFQT decile of the 1980 sample were multiplied by the weights to reproduce the 1975 distribution of scores. The weighting is analogous to selecting the 1980 applicants on the basis of their reference test scores.

TABLE F-1

WEIGHTS FOR SERVICE APPLICANTS IN THE ASVAB 8A CALIBRATION SAMPLE

	Sample		
	Percent in decile		Weight[c]
AFQT Decile	1975[a]	1980[b]	
1. (1-9)	7	9	.8
2. (10-19)	13	25	.5
3. (20-29)	14	14	1.0
4. (30-39)	28	10	2.8
5. (40-49)	14	6	2.3
6. (50-59)	9	9	1.0
7. (60-69)	5	9	.6
8. (70-79)	5	6	.8
9. (80-89)	3	7	.4
10.(90-99)	2	5	.4

[a]Reference test is ACB-73; includes only Army examinees tested with ACB-73 in September through October 1975.
[b]Reference test is ASVAB 5/6/7, correctly scaled.
[c]Weight of 1980 sample to reproduce score distribution of 1975 sample.

The 1980 sample with a weighted distribution of ASVAB 5/6/7 scores is similar to the 1975 sample in three important respects:

- The operational scores, used in enlistment decisions, serve as the reference test.
- The samples have in effect been selected on the basis of the reference test scores.
- Only the experimental tests are given in a separate administration, and examinees may know that the experimental tests do not affect their qualification for enlistment.

Two sources of scale inflation are operating in this analysis. One is prior selection on the reference test, and the second is the use of operational scores as the reference variable. The operational ASVAB 5/6/7 scores in 1980 were relatively free of coaching because the PAFQT (described in appendix G) helped identify recruits who were coached. The effects of prior selection were discussed in the main text. The effects of using operational scores as the reference test should, in general, inflate the scale because examinees typically try harder on an operational test, which affects their qualification for enlistment, than on experimental tests. Also, the operational scores may themselves be inflated by test compromise, which in turn inflates the scale.

The cumulative frequency distributions of the ASVAB 5/6/7 percentile scores and ASVAB 8A raw scores are shown in table F-2. The conversion of ASVAB 8A raw scores to percentile scores is shown in table F-3 and graphed in figure F-1. Both conversions, using ASVAB 5/6/7 and AFQT 7A as the reference, are shown in table F-3.

The scale referenced to ASVAB 5/6/7 is inflated compared to the correct scale referenced to AFQT 7A. The inflation starts at an ASVAB 8A raw score of 33 (percentile score of 8), and reaches a peak at raw scores around 56 (56 converts to a percentile score of 28 for the inflated scale and 21 for the correct scale). Above an ASVAB 8A raw score of 70 the two scales are similar.

The maximum difference between the score scale based on the weighted ASVAB 8A sample and the correct scale is seven percentile points. This difference is much smaller than the maximum difference, 23 points, found between the original and correct scale for ASVAB 5/6/7. Also, the effects of weighting in the simulated ASVAB 8A sample tended to balance each other in the midrange; the conversions were approximately correct in AFQT category IIIA (percentile scores 50-64) and up through percentile scores in the 70s. The original ASVAB 5/6/7 scale, even for Army examinees, remained inflated throughout the category III range.

This analysis confirmed that prior selection on an operational reference test produces an inflated scale. The degree of inflation, however, is about one third that found in the original ASVAB 5/6/7 scale. Apparently other factors in the 1975 sample had a greater effect on the score scale.

We also determined the effects of using operational scores as the reference variable in the full-range, unweighted sample of applicants. These effects are independent of those from prior selection of examinees

TABLE F-2

CUMULATIVE FREQUENCY DISTRIBUTIONS OF WEIGHTED
ASVAB 5/6/7 PERCENTILE SCORES AND ASVAB 8A RAW SCORES

ASVAB 5/6/7[a]				ASVAB 8A[b]					
Percentile score	Cumulative percent	Percentile score	Cumulative percent	Raw score	Cumulative percent	Raw score	Cumulative percent	Raw score	Cumulative percent
1	1	38	62	0-22	2	50	23	78	81
2	2	41	67	23	2	51	25	79	84
3	3	44	72	24	2	52	26	80	86
4	3	47	76	25	3	53	28	81	87
5	3	50	78	26	3	54	29	82	89
6	4	53	81	27	3	55	31	83	90
7	5	56	83	28	4	56	32	84	91
9	7	58	85	29	4	57	34	85	91
10	8	60	87	30	5	58	37	86	92
11	9	62	88	31	5	59	39	87	93
12	10	65	89	32	6	60	41	88	93
13	11	67	90	33	7	61	42	89	94
14	12	70	92	34	7	62	45	90	95
15	14	72	93	35	8	63	48	91	95
16	15	75	94	36	9	64	50	92	96
17	16	77	95	37	9	65	52	93	97
18	18	80	96	38	10	66	55	94	97
19	19	82	96	39	11	67	57	95	98
21	23	84	97	40	12	68	60	96	98

TABLE F-2 (Cont'd)

ASVAB 5/6/7[a]				ASVAB 8A[b]					
Percentile score	Cumulative percent	Percentile score	Cumulative percent	Raw score	Cumulative percent	Raw score	Cumulative percent	Raw score	Cumulative percent
23	25	86	97	41	13	69	63	97	99
25	28	87	98	42	14	70	65	98	99
27	31	89	98	43	15	71	68	99	99
29	33	91	99	44	16	72	70	100-105	99
31	41	93	99	45	17	73	72		
33	49	95	99	46	18	74	74		
35	56	97	100	47	20	75	75		
		98		48	21	76	77		
		99		49	22	77	79		

[a]Distribution of ASVAB 5/6/7 scores explicitly weighted to simulate distribution of ACB-73 of Army examinees tested in September through October 1975.
[b]Distribution of ASVAB 8A scores simulates distribution of ASVAB 5/6/7 scores of the Army examinees in the 1975 sample.

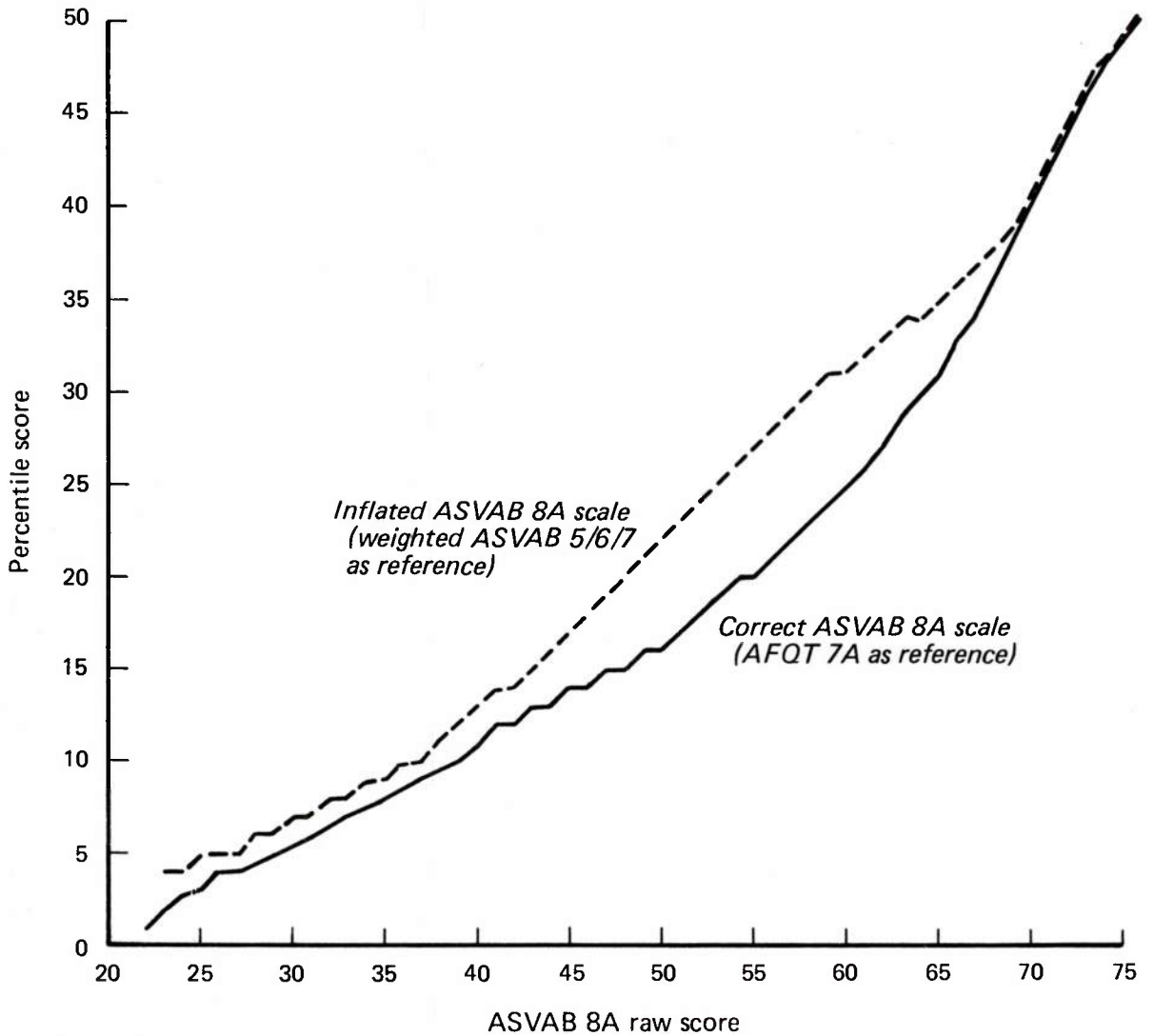

FIG. F-1: ESTIMATED EFFECTS ON ASVAB 8A SCORE SCALE OF SIMULATED SELECTION ON OPERATIONAL REFERENCE TEST

TABLE F-3

ASVAB 8 CALIBRATED TO WEIGHTED[a]
OPERATIONAL ASVAB 5/6/7 SCORES
(Correct scale)

ASVAB 8A raw score	Percentile score: Weighted[a] sample	Percentile score: Correct[b]	ASVAB 8A raw score	Percentile score: Weighted[a] sample	Percentile score: Correct[b]
1-20			54	26	20
21			55	27	20
22		1	56	28	21
23	4	2	57	29	22
24	4	3	58	30	23
25	5	3	59	31	24
26	5	4	60	31	25
27	5	4	61	32	26
28	6	5	62	33	27
29	6	5	63	34	29
30	7	6	64	34	30
31	7	6	65	35	31
32	8	7	66	36	33
33	8	7	67	37	34
34	9	8	68	38	36
35	9	8	69	39	38
36	10	9	70	41	40
37	10	9	71	43	42
38	11	10	72	45	44
39	12	10	73	47	46
40	13	11	74	48	48
41	14	12	75	49	49
42	14	12	76	50	50
43	15	13	77	52	52
44	16	13	78	54	54
45	17	14	79	57	56
46	18	14	80	59	58
47	19	15	81	61	59
48	20	15	82	64	61
49	21	16	83	67	63
50	22	16	84	69	65
51	23	17	85	71	66
52	24	18	86	72	68
53	25	19	87	74	70
88	76	72	97	92	87
89	77	74	98	94	88
90	78	76	99	96	90
91	79	78	100	98	91
92	81	80	101	99	93
93	83	82	102		95
94	85	83	103		97
95	87	85	104		98
96	89	86	105		99

[a]Distribution of ASVAB 5/6/7 scores weighted to simulate distribution of ACB-73 scores of Army examinees tested in September through October 1975.
[b]AFQT 7A used as reference test.

on the reference test. Use of operational scores as the reference variable does result in an inflated scale. Figure F-2 shows both the correct ASVAB 8A scale, referenced to AFQT 7A, and the inflated scale, referenced to ASVAB 5/6/7. The inflation effects started at the bottom of the scale, became more pronounced at about a percentile score of 40 on the correct scale, and persisted to the top of the scale.

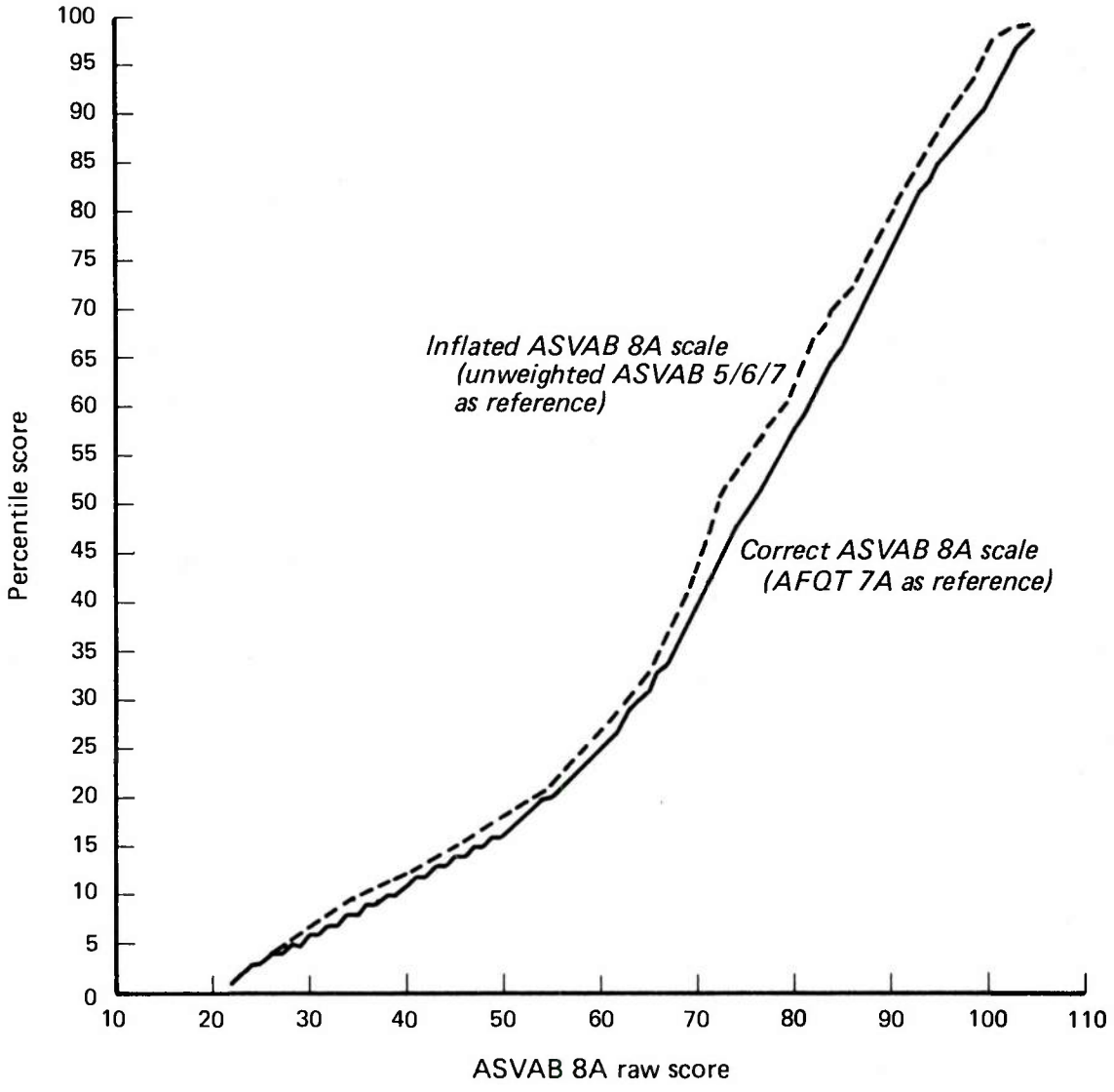

FIG. F-2: EFFECTS ON ASVAB 8A SCORE SCALE OF USING OPERATIONAL SCORES AS THE REFERENCE VARIABLE

REFERENCES

[F-1] CNA, CRC 438, "Normalization of the Armed Services Vocational Aptitude Battery (ASVAB) Forms 8, 9, and 10 Using a Sample of Service Recruits," by William H. Sims and Ann R. Truss, Unclassified, Dec 1980

[F-2] Army Research Institute, Research Report 1301, "Scaling Armed Services Vocational Aptitude Battery (ASVAB) Form 8AX," by Milton H. Maier and Frances C. Grafton, Unclassified, Jan 1981

APPENDIX G

ESTIMATING TEST COMPROMISE USING THE PSEUDO AFQT

APPENDIX G

ESTIMATING TEST COMPROMISE USING THE PSEUDO AFQT

INTRODUCTION

The Pseudo AFQT (PAFQT) was developed by William H. Sims, CNA, to help identify test compromise and recruiter malpractice [G-1]. Because the AFQT is the first hurdle an applicant for enlistment must pass, coaching on the ASVAB to improve test scores is usually focused on the AFQT subtests (Word Knowledge, Arithmetic Reasoning, and Spatial Perception in ASVAB 6/7). A new composite highly correlated with the AFQT, but less likely to be compromised, was developed by Sims. The subtests in the new composite, called the pseudo AFQT (PAFQT), are General Science and General Information (corresponding to Word Knowledge), Mathematics Knowledge (corresponding to Arithmetic Reasoning), and Mechanical Comprehension (corresponding to Spatial Perception). The correlation between the AFQT and PAFQT for ASVAB 5/6/7 is .87 in a sample with no compromise.

In operational use, PAFQT scores are routinely compared to the AFQT scores, and if the difference exceeds a specified value, then the person is retested with another AFQT. The second AFQT score, whether higher or lower, becomes the official score of record.

The purpose of using PAFQT in this report is to compare the PAFQT score distribution with that of the AFQT. The difference between the percentages of examinees in each AFQT category is indicative of test compromise. For the full range of scores, the equipercentile equating technique is the appropriate procedure for putting PAFQT scores on the same scale as the AFQT. Briefly, the equipercentile equating technique sets scores on one variable equal to scores on the other variable that have the same cumulative frequencies in the same or comparable samples. See appendix B for a more complete description of equipercentile equating.

Computing a Pseudo AFQT Score for ACB-73

ACB-73 did not have a PAFQT score; we had to develop one for our analysis. We used the ACB-73 calibration sample to determine the subtests to include in the PAFQT and to calibrate PAFQT to AFQT. Subtests in the PAFQT for ACB-73 are General Science and General Information, corresponding to Word Knowledge in the AFQT; Mathematics Knowledge, corresponding to Arithmetic Reasoning; and Mechanical Comprehension corresponding to Space Perception. The correlation between PAFQT and AFQT was .87 in the ACB-73 calibration sample, which means that PAFQT is a satisfactory estimate of AFQT. The conversion from PAFQT to AFQT raw and percentile scores for ACB-73 is shown in table G-1.

TABLE G-1

CONVERTING PSEUDO AFQT SCORES TO
AFQT SCORES FOR ACB-73

Pseudo AFQT score	AFQT score		Pseudo AFQT score	AFQT score	
	Raw	Percentile		Raw	Percentile
0-19	15	7	50	39	65
20	16	8	51	40	68
21	17	9	52	41	71
22	17	9	53	42	73
23	18	10	54	43	75
24	19	12	55	43	75
25	19	12	56	44	78
26	20	14	57	45	80
27	21	16	58	45	80
28	22	18	59	46	82
29	23	21	60	47	84
30	23	21	61	47	84
31	24	25	62	48	86
32	25	28	63	49	88
33	26	31	64	49	88
34	26	31	65	50	90
35	27	33	66	51	92
36	28	36	67	51	92
37	29	38	68	52	93
38	30	41	69	53	94
39	30	41	70	53	94
40	31	44	71	54	95
41	32	47	72	55	96
42	33	50	73	55	96
43	34	53	74	56	97
44	35	56	75	57	98
45	35	56	76	58	99
46	36	59	77	58	99
47	37	61	78	59	99
48	38	63	79	60	99
49	39	65	80	60	99

Procedures for Estimating the Amount of Compromise

In these paragraphs we explain the rationale for estimating compromise from the difference between AFQT and PAFQT scores and how to compute the percentage of compromise in a sample.

The first consideration is that all aptitude tests have some random error of measurement; that is, they are not perfectly reliable measures of what they purport to measure. The difference between AFQT and PAFQT scores, therefore, has an error component, which we call E. In a representative sample with no compromise or other systematic source of bias in the test scores, half the differences would be positive and half would be negative ($fE_+ = fE_-$, where f is the frequency).

The second consideration is that compromise is assumed to work in only one direction. It always inflates the AFQT score relative to the PAFQT. When PAFQT is subtracted from AFQT, the positive differences (D_+) contain two components: error and compromise ($D_+ = E_+ + C$). Negative differences (D_-), where AFQT is less than PAFQT, contain only an error component ($D_- = E_-$).

In a sample with some compromise, the frequency of negative differences can be used to estimate compromise. For that portion of the sample that was not coached on the AFQT, $fE_+ = fE_-$. The number of examinees with compromised AFQT scores then is simply $fD_+ - fD_-$, or $f(E_+ + C) - fE_- = fC$.

To convert the frequency of compromised cases to percentages, divide by the number of differences observed. That is, the percent of the sample with compromised scores equals $f(D_+ - D_-)/f(D_+ + D_-)$. As an example, assume that 50 percent of the sample has compromised AFQT scores. All 50 percent of these cases would have positive differences; the remaining 50 percent would be split evenly between positive and negative differences; fD_- = 25 percent, $fE_- = fE_+$ = 25 percent, and fD_+ = 75 percent. The percent compromise is computed as $(75 - 25)/(75 + 25) = 50$ percent.

In our estimate of compromise, we are interested only in compromise that moves examinees across AFQT categories. Both the numerator and denominator in that case are reduced, because cases who remain in the same category have a difference score of zero. The formulas developed above, however, still apply.

Estimated Amount of Compromise in the Army Sample

Table G-2 shows the joint distribution of AFQT and PAFQT scores for the 488 Army examinees tested in fall 1975 as part of ASVAB 5/6/7 calibration sample. Because the number of cases at the extremes is small, we collapsed categories I and II and categories IVB, IVC, and V. Also shown in table G-2 is the same joint distribution for the

ACB-73 calibration sample, which had no test compromise. Comparison of the distribution shows that the 1975 sample had relatively more cases whose AFQT scores were higher than their PAFQT scores.

TABLE G-2

JOINT DISTRIBUTION OF AFQT AND PAFQT SCORES

AFQT category	Sample	Percent in AFQT category I&II	IIIA	IIIB	IVA	IVB,IVC,V	Percent of total
I&II	1975[a]	72[c]	15	12	0	1	20.9
	ACB-73[b]	82	13	4	1	0	42.1
IIIA	1975	21	32	35	4	8	13.5
	ACB-73	30	36	30	3	1	21.5
IIIB	1975	6	15	43	17	19	40.4
	ACB-73	6	24	48	13	9	22.0
IVA	1975	0	8	23	27	42	12.7
	ACB-73	1	12	42	18	27	6.9
IVB,IVC,&V	1975	0	2	21	15	62	12.5
	ACB-73	0	4	30	20	46	7.7
Percent of total	1975	20.1	15.0	30.1	12.7	22.1	
	ACB-73	42.3	19.6	24.1	6.5	7.6	

[a]N = 488 Army examinees tested in 1975; PAFQT scores available.
[b]N = 3,815 Army recruits in sample to calibrate ACB-73.
[c]The percent in each row sums to 100.

Table G-3 presents the frequencies in each AFQT category and the estimated amount of compromise in both calibration samples: the ACB-73 sample, where the estimated compromise is 2 percent; the ASVAB 5/6/7 sample, where the estimated compromise is 26 percent. The number of cases in each AFQT category, whose AFQT scores exceed their PAFQT scores cannot be interpreted as direct estimates of the amount of compromise in that category. Although the positive and negative differences are equal in the full-range sample with no compromise, the differences are not equal at the top and bottom of the scale. As seen for table G-3 in the

ACB-73 calibration sample, there appears to be a large amount of compromise in categories I and II (291 cases with positive differences), and a large amount of negative compromise in categories IVB, IVC, and V (156 cases with negative differences). Both values are a function of regression toward the mean of the distribution.

TABLE G-3

ESTIMATED AMOUNT OF COMPROMISE IN ARMY SAMPLE

AFQT category	ACB-73[a] calibration		1975[b] Army examinees	
	fD_+[c]	fD_-[d]	fD_+	fD_-
I&II	291	0	37	0
IIIA	286	239	31	14
IIIB	185	250	71	41
IVA	71	143	26	19
IVB, IVC, & V	0	156	0	23
Total	833	788	165	97
Percent of total sample	22	21	34	20
Percent compromise[e]	2		26	

[a] N = 3,815 Army recruits in ACB-73 calibration sample.
[b] N = 488 Army examinees tested in fall 1975.
[c] D_+ = frequency of AFQT > PAFQT.
[d] D_- = frequency of AFQT < PAFQT.
[e] Percent of compromise in sample = $\frac{fD_+ - fD_-}{fD_+ + fD_-}$.

Table G-2 shows that both the AFQT and PAFQT place about the same percentage of the 1975 sample into AFQT categories I, II, and IIIA (34.4 for the AFQT and 35.1 for the PAFQT). The AFQT placed substantially fewer in categories IVB, IVC, and V than did the PAFQT (12.5 percent versus 22.1 percent). Category IVA was the same for both measures (12.7 percent). Category IIIB contained a larger percentage (40.4 percent) of AFQT scores than of PAFQT scores (30.1 percent). At the low end of the scale, the net effect was to shift examinees from categories

IVB, IVC, and V into category IIIB. The movement in and out of category IVA balanced each other. The amount of compromise in category IIIB could not be easily estimated because of the high concentration of scores in this category.

Our best estimate of the AFQT score distribution for the Army examinees is obtained from the PAFQT scores. Accordingly, we adjusted the frequency of AFQT scores in categories IIIB and IVB, IVC and V. We increased the percentage of Army examinees in the 1975 calibration sample with AFQT scores below 21 by 70 percent, and reduced the percentage in category IIIB correspondingly. The original and adjusted cumulative frequencies of AFQT scores for the Army sample are shown in table G-4. The estimated amount of inflation in the original ASVAB 5/6/7 scale is based on the adjusted ACB-73 scores.

TABLE G-4

CUMULATIVE FREQUENCY OF ACB-73 REFERENCE TEST SCORES IN 1975 ARMY SAMPLE

ACB-73 percentile score	Cumulative percent		ACB-73 percentile score	Cumulative percent	
	Original[a]	Adjusted for compromise[b]		Original[a]	Adjusted for compromise[b]
1-5	1	2	50	79	79
6	3	5	53	81	81
7	4	7	56	83	83
8	5	8	59	85	85
9	7	12	61	86	86
10	9	15	63	88	88
12	12	20	65	89	89
14	15	25	68	90	90
16	18	31	70	91	91
18	20	34	73	93	93
21	25	39	75	94	94
25	30	44	78	95	95
28	34	48	80	96	96
31	41	53	82	97	97
33	47	57	84	97	97
36	54	61	86	98	98
38	60	65	88	98	98
41	65	68	90	99	99
44	71	72	92-99	100	100
47	76	76			

[a]N = 1,151 examinees tested with ACB-73 in September through October 1975.

[b]Frequencies in categories IVB, IVC, and V increased by a factor of .7; category IIIB decreased correspondingly; categories I, II, IIIA, and IVA unchanged.

REFERENCES

[1] CNA, Study 1116, "An Analysis of the Normalization and Verification of the Armed Services Vocational Aptitude Battery (ASVAB) Forms 6 and 7," by William H. Sims, Unclassified, Apr 1978

APPENDIX H

CALIBRATING ASVAB 8/9/10

APPENDIX H

CALIBRATING ASVAB 8/9/10

Three independent studies were designed by the ASVAB Working Group to calibrate ASVAB 8/9/10. The design specified that only one reference test, AFQT 7A, would be used. AFQT 7A was also used to correct the calibration of ASVAB 5/6/7. The reference test, AFQT 7A, and the new ASVAB, were administered in counterbalanced order. Three samples were specified: applicants for enlistment; new recruits from all services; and high school students in grades 11 and 12. Because the reference population contained only males, the calibration samples were also restricted to males. The standard equipercentile equating technique was used in all the studies.

Each sample was analyzed independently by a different research organization. The sample of applicants was analyzed through the combined efforts of the Office of the Secretary of Defense and the Army Research Institute; the sample of recruits by the Center for Naval Analyses; and the sample of high school students by the Educational Testing Service.

The score scale for ASVAB 8/9/10 was based on the combined sample of applicants and recruits. The ASVAB Working Group approved the final results.

Applicant Sample

A nationally representative sample of Armed Forces Examining and Entrance Stations (AFEES) administered AFQT 7A and ASVAB 8 to all applicants for enlistment from 15 January 1980 until the data collection was completed in February 1980. Each AFEES was briefed on the study by a representative of the ASVAB Working Group. Each representative reported that in the sessions observed the AFEES were cooperative and followed good testing practices. Of equal importance was the cooperation of the recruiters in forwarding applicants for testing. On past occasions recruiters were suspected of selectively withholding applicants to avoid experimental testing, or of sending them to places where no extra testing occurred. This problem was minimized in this study because the mobile testing stations manned by military personnel associated with the AFEES were included in the study. The sample should be representative of the applicants processed by the AFEES at that time.

We administered all experimental tests before the operational tests. Fatigue, therefore, should not affect the test scores, and because of the counterbalanced administration of AFQT 7A and ASVAB 8A, motivation should be equal for both the reference test and new test.

As a check on the quality of the test data, we used regression analyses to identify deviant test scores. One analysis was to predict the ASVAB 8A AFQT score from AFQT 7A; the second was to predict the Numerical Operations subtest score from the Arithmetic Reasoning score. Examinees whose scores deviated by more than two standard errors of estimate were deleted from the sample.

The original sample size was 2,620 male applicants. Of this number, 5 percent had deviant AFQT scores. An additional 4 percent had deviant Numerical Operations or Arithmetic Reasoning scores. The final sample of AFEES applicants consisted of 2,375 cases.

Service Recruit Sample

A sample of recruits that represented the current population of new enlisted accessions was used for this analysis. Each service provided its proportional share of the sample (Army--43 percent; Navy--23 percent; Air Force--20 percent; Marine Corps--13 percent). ASVAB 8A and AFQT 7A were administered to 3,799 male recruits from all services. The tests were administered at special sessions conducted by personnel from the reception centers. Each reception center was briefed on the study by a representative of the ASVAB Working Group who also observed at least one testing session.

The editing of the recruit data took a slightly different form than was followed with the applicant sample. The intent was to remove both deviant test sessions and deviant individuals. The first step was to compute mean AFQT scores for each testing session. There were 44 test sessions. A regression analysis was used to identify deviant testing sessions. Sessions that deviated more than 2.5 standard errors of estimate from the regression line were deleted. Nine of the 44 sessions were deviant, and all cases from these sessions were deleted. The second step was to identify individuals with deviant scores. The average regression between AFQT scores was computed, and cases with more than 2.5 standard errors of estimate from the average regression line were deleted. Of the original 3,799 cases, 13 percent were deleted because of faulty testing sessions, and another 3 percent were deleted because of deviant individual AFQT scores. Finally, another 5 percent were deleted because their operational test scores were not available. The final recruit sample had 3,000 cases.

An additional factor that might affect the calibration is the racial-ethnic mix of the sample. The recruit sample was weighted to represent the assumed 1959 racial-ethnic mix, when AFQT 7A was originally calibrated. The assumed mix was 82 percent white, 12 percent black, and 6 percent other. (The applicant sample, even though about one-third of the applicants were black and about 10 percent were hispanic, was not weighted for racial-ethnic mix.)

High School Students

The Educational Testing Service requested schools throughout the country that had participated in the ASVAB High School Testing Program to administer the experimental tests. Of the 180 schools contacted, 40 agreed to participate. The editing of the high school data deleted 9 percent of the cases who attempted very few items on one or more tests. Another 1 percent were deleted because their answer sheets were lost or mutilated, or because of a testing irregularity. All female students were deleted. This left 1,745 usable male cases.

CALIBRATION RESULTS

The conversions from AFQT 8A raw score to percentile score in the three studies are shown in figure H-1. The conversion lines are similar in the bottom end of the scale. There is a tendency for the high school sample to fall to the right of the two military samples. This means that a higher AFQT 8A raw score is required in the high school sample to convert to a given percentile score. The high school sample starts deviating more markedly at about the 20th percentile score, and then becomes more similar again at about the 75th percentile score. The applicant and recruit samples are similar throughout the scale.

In all three studies we found that editing the data had little effect on the score scale. Similarly, weighting the recruit sample to obtain the racial-ethnic mix assumed for the 1959 sample had little effect on the scale. Another finding, supported by other research studies, is that prior selection of the recruits has little effect on the scale. The only consistent difference is that conversion based on high school students result in somewhat lower scaled scores than those based on military samples. A reasonable explanation is that high school students are more literate than school dropouts, but are relatively less superior on nonverbal tests, especially tool knowledge. Because the AFQT from ASVAB 8A has a large literacy component (verbal, arithmetic reasoning, and numerical operations), high school students score higher on this AFQT than on AFQT 7A (word knowledge, arithmetic reasoning, space perception, and tool knowlege); whereas military samples, which contain large percentages of school dropouts, would tend to score relatively better on AFQT 7A. Thus, the results are not surprising.

Based on the similarity of the results for the applicants and recruits, we combined the two samples to construct the final ASVAB 8/9/10 score scale. The cumlative frequency distributions of the AFQT 7A percentile scores and AFQT 8A raw scores are shown in figure H-2. The combined sample of 5,375 cases contained more cases at both extremes than either the applicant or recruit sample, and therefore, should result in more reliable conversions in categories I and V.

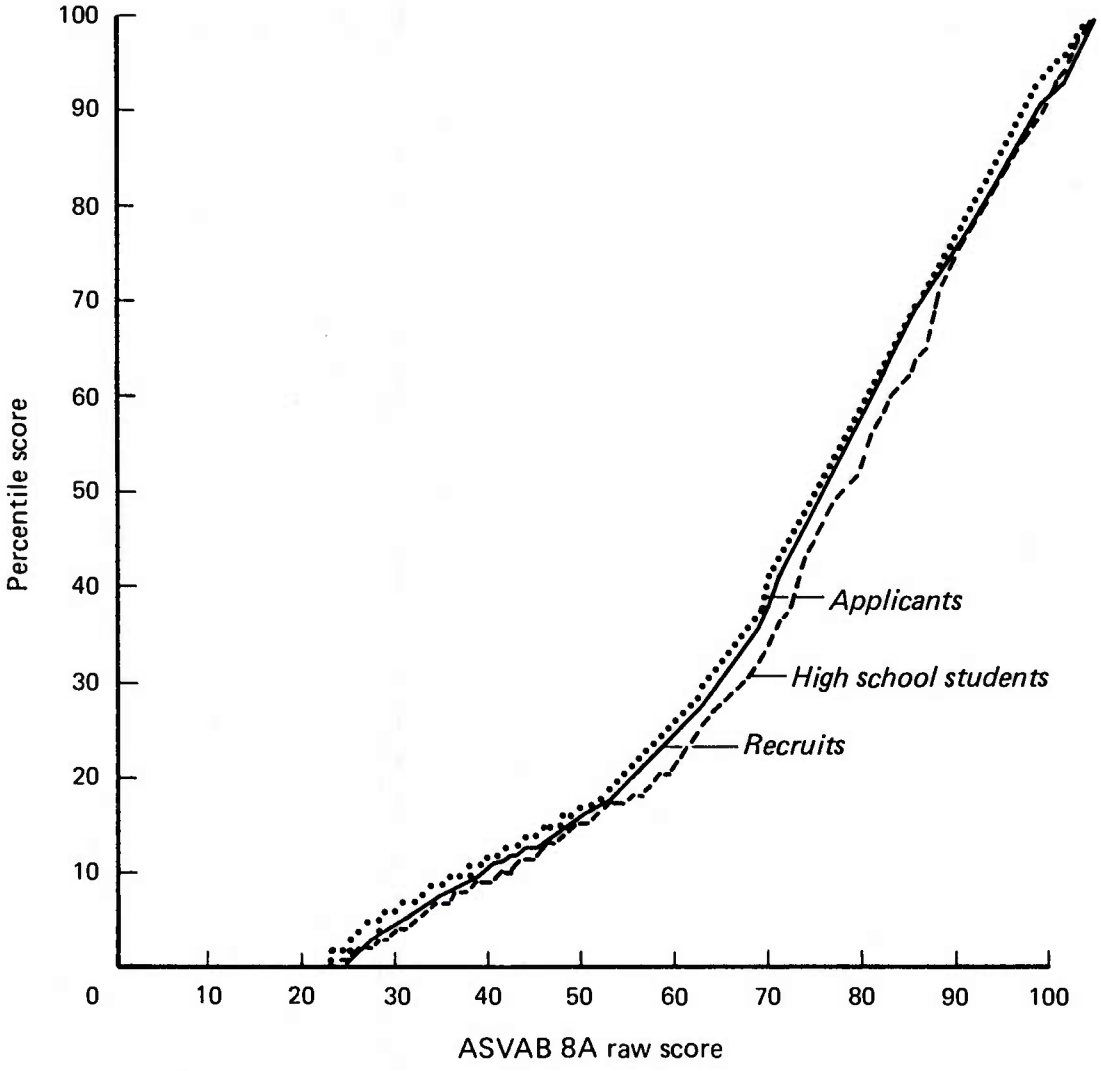

FIG. H-1: CALIBRATION OF ASVAB 8A IN THREE INDEPENDENT SAMPLES

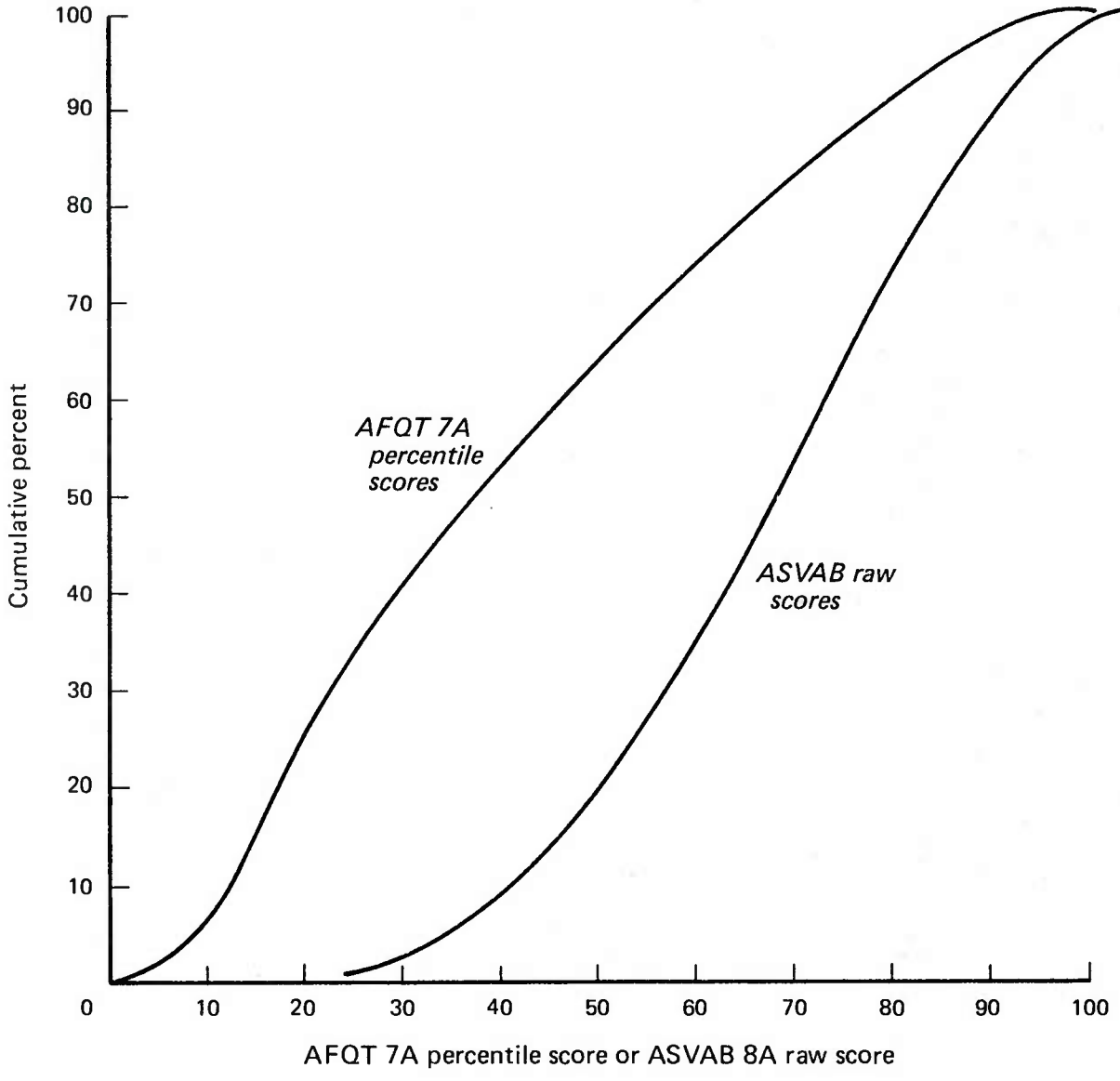

FIG. H-2: CUMULATIVE FREQUENCY DISTRIBUTION OF AFQT 7A AND ASVAB 8A SCORES IN COMBINED SAMPLE OF RECRUITS AND APPLICANTS

A word about smoothing is in order before presenting the final conversions. The two lines in figure H-2 were drawn subjectively rather than analytically. We attempted to use moving averages in the smoothing; because of missing raw scores that arise from formula scoring of AFQT 7A, the smoothed line did not seem appropriate. A similar problem would arise from attempting to average percentile scores. We decided to smooth out some of the bumps in the AFQT 7A distribution, but as can be seen, we generally stayed close to the data. A second smoothing occurred after reading the converted values from the two lines. An operational requirement is that percentile scores occur at breakpoints in the grouping of AFQT scores (percentile scores of 10, 16, 21, 31, 50, 65, and 93). The final smoothing accomplished this.

The final conversion, proposed by the ASVAB Working Group and adopted for operational use, is shown in figure H-3. The conversion shows the desired properties:

- Discrimination at the low end, in categories IV and V, should be reliable; one or two raw scores correspond to each percentile score
- Discrimination throughout the score range appears to be adequate
- The progression in percentile scores is orderly.

CONTRAST BETWEEN CALIBRATION OF ASVAB 5/6/7 AND ASVAB 8/9/10

The error in the ASVAB 5/6/7 score scale led to increased emphasis on ensuring that ASVAB 8/9/10 was scaled correctly. Whereas compromises with operational requirements were made in 1975 when ASVAB 5/6/7 was calibrated, the procedures for ASVAB 8/9/10 followed good testing practice. The contrast is shown in table H-1.

Reference Test

Two reference tests were used to calibrate ASVAB 5/6/7. One was ASVAB 2, and the other was ACB-73. Because ASVAB 2 was not an operational test, these scores could not be affected by coaching. The ACB-73 scores, however, were the operational scores of record, subject to coaching by recruiters. The tests had been in continuous operational use for more than 2 years when they were used as the reference measure. ASVAB 8/9/10 by contrast used a single reference test, AFQT 7A. This test had not been used operationally for more than 5 years, and there is no reason to question the accuracy of the scores.

Samples

Two samples--recruits and applicants, representing different ends of the score scale--were combined to scale ASVAB 5/6/7. Air Force and

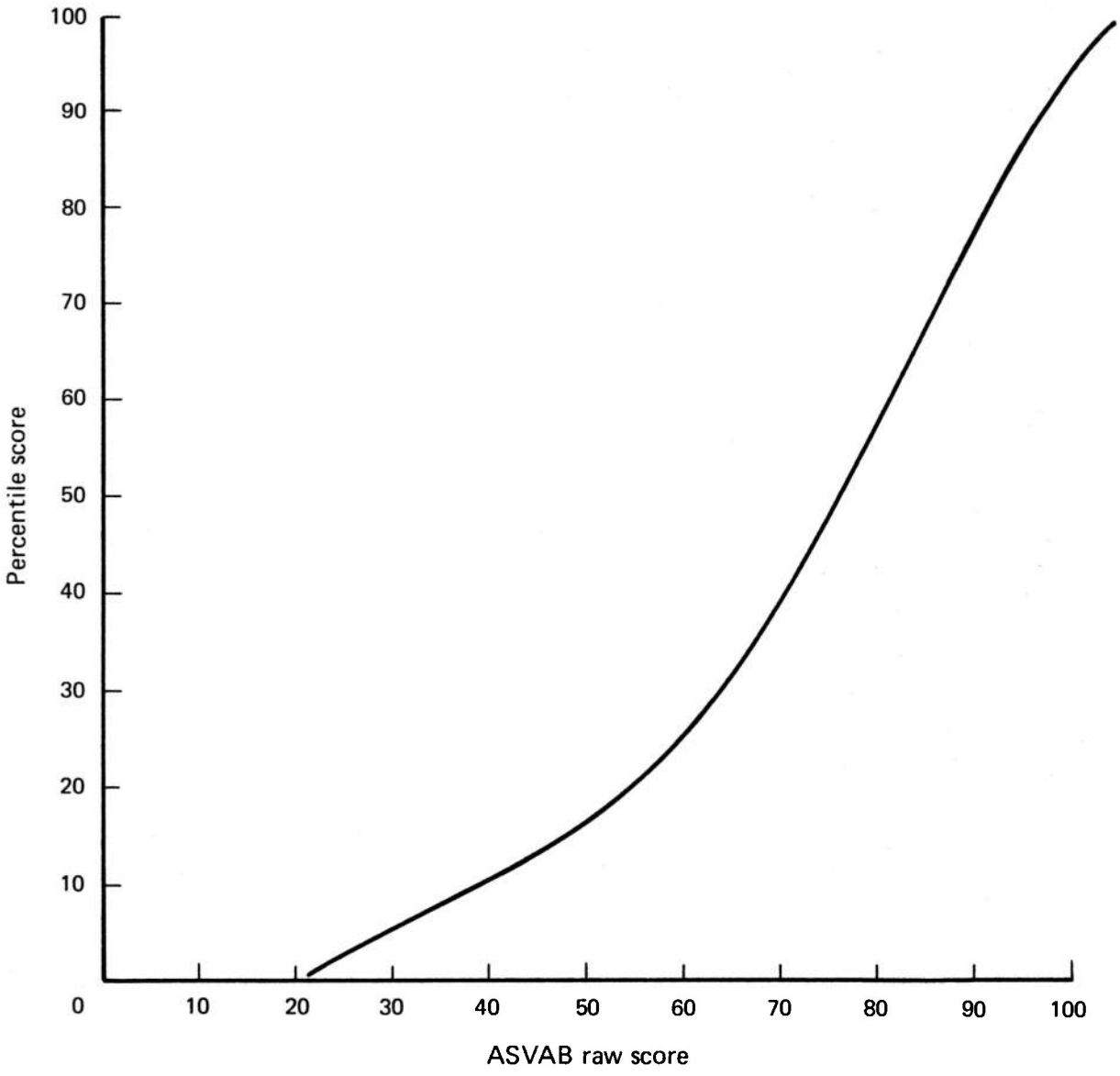

FIG. H-3: FINAL CALIBRATION OF ASVAB BASED ON COMBINED SAMPLE OF RECRUITS AND APPLICANTS

Navy recruits were used to set the top half, and Army applicants the bottom half. No separate analyses of the recruits and applicants was performed prior to constructing the final ASVAB 5/6/7 scale. In contrast, three independent samples were used to calibrate ASVAB 8/9/10. Not only were the samples independent, but the analysis of each was also conducted independently. All three studies used the equipercentile equating technique, but followed different procedures in editing the data and in the smoothing. The consistency of the three sets of results lends credence to their accuracy.

TABLE H-1

COMPARISON OF ASVAB 5/6/7 AND ASVAB 8/9/10 CALIBRATION

	ASVAB 5/6/7	ASVAB 8/9/10
• Reference Test	ASVAB 2 operational ACB-73	AFQT 7A
• Samples	One	Three independent
• Testing conditions at AFEES	Obvious difference between reference and experimental tests	No difference between reference and experimental tests

Testing Conditions

Perhaps the most salient difference is in testing conditions at AFEES. In 1975 the AFEES were supposed to administer ACB-73 and the experimental ASVAB 5/6/7 in counterbalanced order, but there is strong reason to question whether this was actually accomplished. In contrast, the administration of ASVAB 8/9/10 and AFQT 7A was counterbalanced, and both preceded operational testing. The test-taking behavior should be comparable on both of the experimental measures.

The evidence indicates the ASVAB score scale has been restored to its original meaning.

APPENDIX I

ACCURACY OF SCORING ASVAB 5/6/7

APPENDIX I

ACCURACY OF SCORING ASVAB 5/6/7

The transition from one test battery to another invariably disrupts established practice, and people need time to adjust. Persons with scores from the older tests are still reported in the distributions for the period when the new test is implemented. For example, applicants tested in December 1975, but not enlisting until January 1976, would have been administered the old test, but may be included in the January 1976 score distributions. Another factor may be the scoring of the unfamiliar tests; the quality control checks may not have eliminated all the errors. One task therefore is to examine the operational data records to determine which scores are trustworthy. The scores that are judged accurate will be compared to evaluate the effects on score distributions of changing from ACB-73 to ASVAB 5/6/7.

The first step in analyzing the shift or lack of shift in the percentages of Army applicants in each AFQT subcategory is to ascertain which tests were included in these score distributions. In December 1975 the only form administered to Army applicants at the examining stations was ACB-73; in January 1976 and immediately thereafter, forms other than ASVAB 5/6/7 may have been administered. The records of the test forms administered during the first 6 months of 1976 were obtained from the Defense Manpower Data Center (DMDC); the frequencies are shown in table I-1. Most of the applicants were tested with ASVAB 5/6/7. The number of cases with unknown forms was about 2,500 each month. The applicants coded as having taken the Navy Basic Test Battery, forms 6 and 7, probably took ASVAB 5/6/7; although the score distributions of these cases were similar to those who took ASVAB 5/6/7, they were excluded from further analysis. In January 1976, 1,763 of the scores were obtained from ACB-73; the number decreased in each succeeding month, and the numbers shown from ACB-73 probably are an accurate count.

Separate score distributions were obtained for the applicants who were coded as having taken ASVAB forms 6 or 7. The distributions for the two forms were virtually identical, which indicates that the two forms are parallel, and they were pooled. The score distributions for persons who took ASVAB 5/6/7 were compared to the distribution of the total number tested; the comparisons are shown in table I-2. Also shown in table I-2 are the distributions of scores for those coded as taking ACB-73 in October and December 1975 compared to the total tested. In each case, the distributions differ only slightly. The ACB-73 distributions in 1975 are within one-half of 1 percent, except category IVC in October 1975 which differs by 0.6 percent. The ASVAB 5/6/7 distributions in 1976 also agree closely with the total, except in category V. In this case the ASVAB 5/6/7 percentage exceeds the total by about 1 percent in January and July.

TABLE I-1

DISTRIBUTION OF TEST FORMS ADMINISTERED IN 1976
TO ARMY MALE APPLICANTS

Test form	Jan	Feb	Mar	Apr	May	Jun
ASVAB 6	18,387	14,126	14,937	10,111	7,939	9,516
ASVAB 7	13,279	12,545	12,641	9,641	8,693	10,051
ACB-73	1,763	600	533	345	218	62
NBTB 6[a]	0	4	1	512	1,872	1,933
NBTB 7[a]	0	59	488	1,851	2,329	1,933
Unknown and miscellaneous forms	3,378	2,906	3,813	2,932	2,690	3,676
Total tested	36,807	30,240	32,413	25,392	23,741	27,171

[a]NBTB = Navy Basic Test Battery, used for Navy applicants until ASVAB 5/6/7 implemented; these applicants probably took ASVAB 5/6/7.

The data suggest no serious error in the distributions arising from the mixture of test forms, and the score distributions of record appear to be accurate for the applicants tested. Analysis of the scores of record indicates that the score distributions are sufficiently accurate to permit reliable comparisons.

TABLE I-2

SCORE DISTRIBUTIONS FOR ARMY APPLICANTS TESTED WITH ACB-73 OR ASVAB 6/7 COMPARED TO TOTAL NUMBER TESTED

	ACB-73 versus total				ASVAB 6/7 versus total					
	Oct 1975		Dec 1975		Jan 1976		Jul 1976		Sep 1976	
AFQT category	ACB-73	Total	ACB-73	Total	ASVAB	Total	ASVAB	Total	ASVAB	Total
I & II	19.2	19.7	20.4	20.8	27.2	27.6	25.2	26.1	15.8	16.4
IIIA	15.8	16.2	16.0	16.4	10.1	10.7	10.0	10.3	14.1	14.3
IIIB	26.4	26.7	26.2	26.5	21.3	21.9	23.7	24.0	30.5	30.7
IVA	11.5	11.3	11.2	10.9	12.7	12.4	9.7	9.6	9.3	9.2
IVB	6.1	6.1	6.3	6.2	6.3	6.1	6.6	6.5	6.6	6.5
IVC	11.3	10.7	10.8	10.4	6.1	5.8	6.6	6.3	9.1	8.8
V	9.7	9.3	9.2	8.9	16.4	15.5	18.2	17.2	14.6	14.1
Total	25,066	32,042	24,592	28,846	31,703	36,807	21,434	29,650	22,974	28,870

REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS BEFORE COMPLETING FORM

1. REPORT NUMBER: CRC 457

2. GOVT ACCESSION NO.:

3. RECIPIENT'S CATALOG NUMBER:

4. TITLE (and Subtitle): Original Scaling of ASVAB Forms 5/6/7: What Went Wrong

5. TYPE OF REPORT & PERIOD COVERED:

6. PERFORMING ORG. REPORT NUMBER:

7. AUTHOR(s): Milton H. Maier, Ann R. Truss

8. CONTRACT OR GRANT NUMBER(s): N00014-76-C-0001

9. PERFORMING ORGANIZATION NAME AND ADDRESS:
Center for Naval Analyses
2000 No. Beauregard Street
Alexandria, Virginia 22311

10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS:

11. CONTROLLING OFFICE NAME AND ADDRESS:
Deputy Chief of Staff
Headquarters, Marine Corps
Washington, D.C. 20380

12. REPORT DATE: March 1981

13. NUMBER OF PAGES: 153

14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office):

15. SECURITY CLASS. (of this report): Unclassified

15a. DECLASSIFICATION/DOWNGRADING SCHEDULE:

16. DISTRIBUTION STATEMENT (of this Report):

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report):

18. SUPPLEMENTARY NOTES
This Research Contribution does not necessarily represent the opinion of the Commandant, Marine Corps.

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)
AFQT (Armed Forces Qualification Test), Aptitude Tests, ASVAB (Armed Services Vocational Aptitude Battery), Enlisted Personnel, Ratings, Recruiting, Scaling Factors, Test Methods, Validation

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

By April 1976, 4 months after it was introduced, the traditional meaning of scores on the Armed Services Vocational Aptitude Battery, forms 5, 6, and 7 (ASVAB 5/6/7), was being questioned. Scores were found to be too high compared with the traditional reference of the ASVAB score scale in the above-average range. By 1980, scores were also verified to be too high in the below-average range. Our analysis to find the errors in the score scale suggested three reasons:

- Incorrect scoring of the reference test used with the sample of Navy and Air Force recruits

20

- Coaching on the reference test used for Army examinees
- Using operational test scores as the reference variable for Army examinees and excluding those who did not qualify for enlistment from the calibration sample.

The explanations accounted for almost all the inflated scores on the original scale for ASVAB 5/6/7 above a percentile score of 50 and below a percentile score of 15. Between percentile scores of 15 and 50, however, a residual of up to one-third the difference between the original scale and the traditional ASVAB remained unexplained.

On 1 October 1980, a correct score scale for ASVAB 5/6/7, accurately calibrated to the traditional reference, was implemented.

Based on our analysis we conclude that the original 1976 ASVAB 5/6/7 score scale was in error and that the traditional meaning of the ASVAB scores has been restored.